SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.235
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparison of Naïve Bayes Algorithm, C4.5 and Random Forest for Service Classification Ojek Online

Hermanto 1st
Master of Computer Science-Postgraduate Programs
STMIK Nusa Mandiri
Jakarta, Indonesia
Hermanto.hmt@bsi.ac.id

Sandra Jamu Kuryanti 2nd
University Bina Sarana Informatika
Jakarta, Indonesia
Sandra.sjk@bsi.ac.id

Siti Nur Khasanah 3rd
STMIK Nusa Mandiri Jakarta
Jakarta, Indonesia
Siti.skx@nusamandiri.ac.id

***Abstract* —** Online transportation is currently the most popular transportation because of the ease of using this transportation service in mobile phone applications. Some people express their opinions and opinions about users of public transport through social media sites and other websites. This opinion can be used as a material for sentiment analysis to find out whether public transport services are positive or negative. The purpose of this study was to find out the sentiments in the tweet opinion and to find out the results of the classification of the Naive Bayes method, C4.5 and the Random Forest algorithm that were used and compared. In this study, from the results of testing with performance measurement the three algorithms use Cross Validation, Confusion Matrix and ROC Curve. There are differences in the value of accuracy between the algorithms applied. In this study using datasets originating from the @GrabID account and @gojekindonesia. Accuracy produced in this study, resulted in a naïve bayes classification algorithm having an accuracy value of 69.18% and AUC value of ¬0.771 so that included in fair classification, for random forest algorithm has an accuracy of 66.34% and AUC value of 0.738 so that it is included in fair classification, while for the c4.5 algorithm has an accuracy of 65% and an AUC value of 0.686 so that it belongs to poor classification. From these results it can be concluded that the naïve bayes algorithm has a higher accuracy compared to random forest and c4.5 algorithms, so that the difference in accuracy between naïve bayes and random forest is 2.84%, while the difference between naïve bayes and c4.5 is 3 , 53%. So that it can be concluded in this study the algorithm that has the best performance is the Naïve Bayes algorithm.

Keyword - Sentiment Analysis; Naïve Bayes; C4.5 and Random Forest, Online Transportation

I. INTRODUCTION

Transportation is one of the supporting needs of the community which is used as a means to move from one place to another. Along with the development of technology currently influences system performance in all aspects including in the aspect of transportation. Currently, online motorcycle taxi is the latest public transportation trend among the community because of the ease of using this transportation service through application devices without having to conventionally come to conventional places to use this transportation service. Along with the development of online motorcycle taxi services, people often talk about it by giving their opinions and opinions through various media, one of which is Twitter social media. The opinions give by the community to online motorcycle taxi services also vary. On Twitter, online motorcycle taxi companies have official accounts to provide up-to-date information about services and accommodate tweets of comments from the public and customers (Nugroho, dkk, 2016).

Twitter is an online social networking and microblogging service that allows users to send and read text-based messages up to 140 characters, known as Twitter. Twitter was founded in March 2006 by Jack Dorsey, and the social networking site was launched in July. Since its launch, Twitter has become one of the ten most visited sites on the Internet, and has been

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.235
e-ISSN : 2541-2019
p-ISSN : 2541-044X

dubbed short messages from the Internet. On Twitter, unregistered users can only read Twitter, while registered users can write Twitter via the web interface, short message (SMS) interface, or through various mobile device applications.

The opinions given by the public regarding the services provided by these online motorcycle taxi services are varied, such as giving opinions about satisfaction or reduced satisfaction of the people who use these services, so with the many opinions given, making people selective in choosing motorcycle taxi service providers online with these conditions, from the side of the online motorcycle taxi service provider companies can find out community satisfaction with the services provided so they can improve the quality of services that will be provided to the customer continuously.

Sentiment analysis is one part of text mining that is used as an analytical tool to understand, extract and process textual data automatically to obtain sentiment information contained in an opinion sentence (Liu, 2010).

There are several studies related to sentiment analysis conducted by several researchers such as Dey et al (2016) analyzing sentiment for review of film datasets and hotels using Naïve Bayes, C4.5 and Random Florest. (Susanti et al., 2017), research on sentiment analysis of GSM services with the dataset used is derived from twitter and naïve bayes as its classification technique. All algorithm models from the above research are used to analyze sentiment from text.

In this study, the authors used the Naïve Bayes algorithm, C4.5 and Random Forest which will be compared to be applied in the text classification process of opinions on online motorcycle taxi services, the results of which can determine which method has the best accuracy that can be applied in classify tweets with the value of positive or negative sentiment on Twitter.

II. LITERATURE REVIEW

A. *Data Mining*

Data mining itself according to (Han, 2012) is an attempt to find interesting patterns from large amounts of data, which can be stored in databases, data warehouses, or other storage areas. Likewise with data consisting of tweets, the amount of data is abundant and of course it has interesting patterns that can be utilized.

Data mining is a process that employs one or more computer learning techniques (Machine Learning) to analyze and extract knowledge automatically. Another definition of the world is induction-based learning is the process of forming general concept definitions that are carried out by observing specific examples and concepts to be learned (Hermawati, 2013).

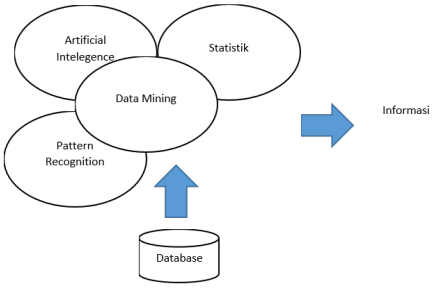


Figure 1. Root of Data Mining Science

There are other terms that have the same meaning as data mining, namely knowledge discovery (KDD). Indeed, data mining or KDD aims to utilize data in the database by processing it to produce useful new information. As illustrated in figure 1.1, if tracked by science roots, it turns out that data mining has four roots in the scientific field as follows:

1. Statistics

This field is the oldest root, without statistics, there is no data mining. Using classical statistics it turns out that processed data can be summarized in what is commonly known as exploratory data analysis (EDA). EDA is useful for identifying systematic relationships between variables/ features when not enough natural information is carried.

2. Artifical intelligence (AI)

This field of science is different from statistics. The theory is built on heuristic techniques so that AI contributes to information processing techniques based on human reasoning models. One branch of AI, namely machine learning or machine learning, is the most important scientific discipline presented in building data mining, using techniques where computer systems run with 'training'.

3. Pattern Recognition

Actually, data mining is also a derivative of the field of pattern recognition, but it only processes data from the database. Data taken from the database to be processed is not in the form of relations, but in the first normal form so that the data set is formed into the first normal form. However, data mining has a characteristic that is the search for association patterns and sequential patterns.

4. Database system

The fourth root of the science field of data mining that provides data similar information that will be 'explored' using the methods mentioned previously

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.235
e-ISSN : 2541-2019
p-ISSN : 2541-044X

including economics, health services and knowledge research (Prasetyo, 2014).

B. *Main Foundation in Data Mining*

Data Mining has four basic foundations (Han, Kamber, and Pei, 2012), namely:

1. Classification
   Classification is a form of data analysis that attempts to extract a model that explains the important classes contained in the data.
2. Clustering
   Clustering is grouping a group of data objects into several groups or clusters so that objects in a cluster have a high resemblance, but these objects have an incompatibility with different clustered objects.
3. Association
   Association is a process to find relationships between data objects that occur repeatedly in a dataset.
4. Outlier Detection
   Detection of installments is a process to find data objects with characteristics that are very different from the characteristics that the data object should have.

C. *Model Process Data Mining*

Data mining is a process, so that the process must be in accordance with procedures in data mining, the most popular data mining process is the Cross-Industry Standard Process for Data Mining (CRISP-DM) process, following the steps of the CRISP-DM process (Putler and Krider, 2015):

1. Business Understanding
   The initial phase in the CRISP-DM process focuses on understanding project objectives and needs from a business perspective, then converts this knowledge into problem definitions and designs initial data mining plans to achieve project goals.
2. Data Understanding
   The data understanding phase begins with collecting initial data such as recognizing data, identifying data quality problems, or for detecting an interesting subset of data
3. Data Preparation
   In the data preparation stage is a final phase in preparing datasets to be used in the construction of models that are built from the raw data available for use.
4. Modeling
   In the modeling phase, it is actually a model stage built and in value. Common tasks related to this stage are: selection of modeling techniques, generating test designs, building models and assessing models.
5. Evaluation
   At this stage a model that has been made that appears to have high quality from the perspective of data analysis has been produced. Before proceeding to the application stage. The main task in this stage is to evaluate the results of the model, review the process, and determine the next stages.
6. Deployment (using models in everyday business)
   At this stage of deployment is applying a model developed into relevant business processes in an organization.

D. *Classification*

Classification is a data analysis process that produces models to describe the classes contained in the data (Han, Kamber, & Pei, 2012). These models are called classifiers. So, this classifier will be used to compile the classes contained in the data. There are many types of classification algorithms, two of which are the Decision Tree and the Nearest Neighbor (k-NN).

Classification can be defined in detail as a job that conducts training/ learning on the target function *f* which maps each vector (feature set) into one of the available numbers of y class labels. The training work will produce a model which is then stored as memory. The model in classification has the same meaning the black box, where there is a model that receives input and is then able to think about the input and provide answers as outputs of the results of his thoughts (Prasetyo, 2014).

Each classification technique uses a learning algorithm to get a model that best meets the relationship between the set of attributes and class labels in the input data. Usually, the input from the classification model is a set of records (training set). Each record includes a set of attributes which one of its attributes is a class. The model for class attributes is a function of other attribute values. A test set is used to determine the accuracy of the model. Usually, the dataset provided is divided into training and test sets, where training sets are used to build models and test sets are used to validate (Hermawati, 2002).

E. *Text Mining*

As one type of Data Mining is a methodology that analyzes textual data that is not easy to process algorithmically, unstructured, but is the most common form of data in the process of information exchange (Witten, 2005).

Text mining (TM) can be defined as a scientific process in which a researcher interacts with a set of documents using various devices to analyze the text

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.235
e-ISSN : 2541-2019
p-ISSN : 2541-044X

contained in the document (Feldman R. & Sanger James, 2007). The main objective of TM is to analyze information to find patterns (Aggarwal & Zhai, 2012).

Text mining is an area of new and interesting research which tries to solve the problem of information overload by using data mining techniques, machine learning, Natural Language Processing (NLP), Information Retrieval (IR), and knowledge management. Text mining involves preprocessing stages of document collections such as text categorization, information extraction, term extraction (Feldman and Sanger, 2007).

F. Twitter

Twitter is micro blogging and social networking site that is popular with a registered user base of around 650 million per year that allows users to send text messages at most 140 characters (tweets) (Wahyudi and Putri, 2016). Twitter is a micro blog where people can send messages in real time about their opinions on various topics, discuss popular issues, submit complaints and express positive or negative sentiments for the products they use in their daily lives. Even some product manufacturing companies learn user reactions to their products via Twitter (Wahyudi & Putri, 2016).

G. Rapid Miner

Rapid Miner is open software. RapidMiner is a solution for analyzing data mining, text mining and predictive analysis. RapidMiner uses a variety of descriptive and predictive techniques to provide insight to users so that they can make the best decisions. RapidMiner has approximately 500 data mining operators, including operators for input, output, data preprocessing and visualization. RapidMiner is stand alone software for data analysis and as a data mining machine that can be integrated into its own products. RapidMiner is written using java language so that it can work on all operating systems (Aprilia et al: 2013).

H. Naïve Bayes

Naive Bayes Classifier algorithm is an algorithm that is used to find the highest probability value to classify test data in the most appropriate category (Feldman and Sanger, 2007).

Naive Bayes is the simplest calculation of the Bayes theorem, because it can reduce computational complexity into a simple multiplication of probabilities. In addition, the Naive Bayes algorithm is also capable of handling data sets that have many attributes (Sartika and Sensuse, 2017).

I. C4.5

Decision tree algorithm or can also be called C4.5 algorithm is an algorithm that has a concept on a divide and conquer approach for a classification process of a problem (Andriani, 2013). The decision tree algorithm in the decision tree formation can be a decision tree C4.5 (Han et al., 2015).

J. Random Forest

Random Forest (RF) is a derivative algorithm or development of a single decision tree. The RF algorithm consists of several trees or decision trees where each tree is trained in the sample data (Sambodo, Rahayu, & Indriasari, 2014).

According to (Nugroho & Emiliyawati, 2017) explaining that the Random Forest Method (RF) is a method that can improve the results of accuracy, because in generating a child node for each node is done randomly.

K. Testing *K-Fold Cross Validation*

Validation is a process to evaluate the accuracy of predictions from a data mining model. K-Fold Cross Validation is a validation technique by dividing data randomly into k parts and each part will be carried out a classification process (Han, Kamber and Pei, 2012).

L. *Confusion Matrix*

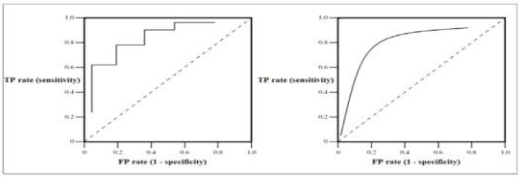
Confusion Matrix is a visualization tool commonly used in supervised learning. Each column in the matrix is an example of a prediction class, while each row represents an actual class event (Gorunescu, 2011).

Table 1. Model Confusion Matrix

Classification		*Predited Class*	
		Class = Yes	*Class = No*
Observed Class	*Class = Yes*	A (*true positive-tp*)	B (*false negative-fn*)
	Class = No	C (*false positive-fp*)	D (*true negative-tn*)

M. *Kurve ROC*

The ROC curve can be used to test and assess the results of visual classification performance and those used to express confusion matrix. The ROC curve is a two-dimensional graph with false positive as a horizontal line and true positive as a vertical line (Vecellis, 2009).



Source: Gorunescu (2011)

Figure 2. Grafik ROC

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.235
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The level of accuracy can be diagnosed as follows (**Gournescu, 2011):**
Accuracy 0.90 – 1.00 = *Excellent classification*
Accuracy 0.80 – 0.90 = *Good classification*
Accuracy 0.70 – 0.80 = *Fair classification*
Accuracy 0.60 – 0.70 = *Poor classification*
Accuracy 0.50 – 0.60 = *Failure*

N. *Framework*

In completing this study, the author made a framework that was used as a reference in this study so that research can be done well. This research consists of several stages, namely: Problem, Approach (Naïve Bayes, C4.5 and Random Forest, Development (Rapidminer), Implementation (Tweets the Indonesian language regarding online motorcycle taxi services, Experiments with Model CRISP-DM), Measurement (Confusion Matrix, ROC Curve), Result.

The problem in this study is that the exact method with the best accuracy is not yet known for the classification of opinions that have a negative sentiment and positive sentiment towards online motorcycle taxi services using the method used is the Naïve Bayes algorithm, C4.5 and Random Forest.

III. PROPOSED METHOD

The research method used in this experiment uses a standard methodology in Data Mining research, which is a Cross-Standard Industry for Data Mining (CRISP-DM) model consisting of 6 phases (Putler and Krider, 2015), namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment.

A. Business understanding

In the business understanding stage, it can be referred to as the stage of understanding research, determining the purpose of research projects in formulating defining data mining problems. The more people who use social media, the more data in the form of opinions are generated, in those opinions not only talking about one topic that is conveyed.

This study aims to create a classification system to determine the assessment of online motorcycle taxi services originating from two services used by the public through public opinion delivered on social media such as twitter and implementing classification methods in text mining.

B. Data Understanding

In the Data Understanding stage, data collection is carried out, an analysis of data investigations from datasets that have been collected through the crawling process using the twitter plug in crawler contained in rapidminer tools connected with twitter using OAuth twitter access assistance, data collected in the process of crawling data tweets this is then saved in the excel file format. The main data sources used in this study use the twitter dataset of public opinions regarding online motorcycle taxi services delivered on social media twitter, the datasets collected in this study are based on certain queries that have been determined based on @GrabID and @gojekindonesia queries. Based on the dataset that has been collected which then continues to the next stage, namely by giving a label to the dataset.

C. Data Preparation

In the stage of data preparation, it is a stage to prepare data to be applied in modeling, which previously came from the initial raw data to the classification stage. In this stage is a process known as pre-processing is a stage that contains many activities as follows (Nugroho, et al, 2016):

1. Case Folding or Transform Case
   In the folding case process this is a process of uniforming the shape of letters and the elimination of punctuation. In this case it only accepts Latin letters between a to z.
2. Tokenizing
   Tokenizing is a process for separating words, each word will be separated based on the spaces found.
3. Stemming
   Stemming is a process of changing the word affix into a basic word.
4. Filtering
   The filtering process is a process for selecting Indonesian-language tweets and the process of removing non-essential words from the results of tokens.

D. MODELING

The next stage in the CRISP-DM model is modeling where this stage directly uses data mining techniques. At this stage, the dataset that was made in the previous stage is used as input for the classification algorithm, which is used as a training dataset. In this study two types of algorithms will be used, namely, Naïve Bayes, C4.5 and Random Forest

E. Evaluation

At the evaluation stage, it is a stage in classification because at this stage testing will be determined for accuracy. The testing phase is to see the results of accuracy in the classification process of Naïve Bayes Algorithm, C4.5 and Random Forest Algorithm and evaluation with ROC Curve. The Evaluation aims to determine the usefulness of the

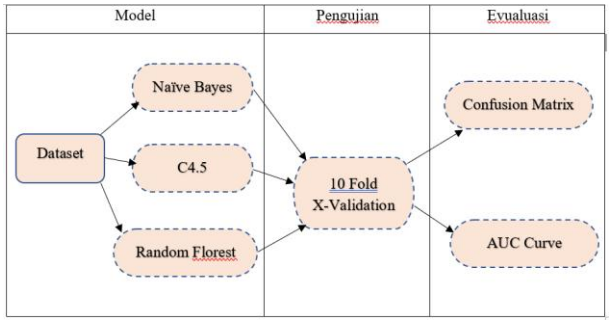
SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.235
e-ISSN : 2541-2019
p-ISSN : 2541-044X

model that we have succeeded in making in the previous step.

F. Deployment

At the stage of deployment, is the time when the results of all the previous stages are used in real terms. Based on the research conducted, a new pattern, information and knowledge has been produced in the data mining process

III. PROPOSED METHOD



Source: Research result (2017)

Figure 3. The proposed model

IV. RESULT AND DISCUSSION

A. Implementation of the Methodology

Based on the research methodology described in chapter III, the following methodology implementation was carried out in this study.

B. Data Collection

In this study the author uses data tweets as the object of this research. The data used in this study through the process of collecting data from social media twitter by crawling or pulling twitter data from several online motorcycle taxi service account accounts, in this study the data used is user tweets originating from @GrabID and @ accounts. gojekindonesia as the data source.

Based on the description of the process of data collection the above research was carried out by utilizing rapidminer tools that are connected with TwitterOAuth libraries and the total tweets collected in the process of data collection were successfully obtained as many as 7,702 tweets. From all the data tweets are then stored in the excel file. From the total tweets, the tweets that have duplicated text tweets to be removed are removed with a total of 2688 tweets.

C. Labeling Data (Labeling)

The data labeling process is done by providing prediction values of predetermined class sentiments, namely positive and negative. The labeling process in this research is done manually to provide classification values by calculating the many texts contained in each document, the labeling process is carried out as follows (Hadi et al., 2017):

1. Determine words that have positive meanings too
2. negative.
3. Calculate the number of positive and negative words in the document.
4. If the number of positive words a number of negative words, then the label of the sentiment is positive.
5. If the number of positive words is <negative number, then the sentiment label is negative.

The results of this tweet labeling process are in the form of a corpus which will be continued in the pre-processing stage. The following is the distribution of the amount of twitter data that has been labeled:

Table 1. The Dataset has been labeled

No	Label	
		Positif
1	Positif	1085
2	Negatif	1603
Total		2688

Source: Research result (2018)

A. Data *Preparation*

After all the data already has a class then the next step is to pre process the corpus that has been produced in the previous stage. The pre-processing stage is divided into several steps, including:

1. *Case Folding*

At the stage of folding case, it is a step to change all letters in a tweet to lowercase.

2. *Tokenization*

Tokenization is a process that is used to parse text into its constituent word units based on word separators in sentences, such as spaces. But before tokenization every tweet is separated from special characters and certain objects so that tweets become cleaner to eliminate unnecessary tags that usually appear on tweets, such as links starting with http, mentioned beginning with @ symbol, the hashtag starts

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.235
e-ISSN : 2541-2019
p-ISSN : 2541-044X

with the # symbol, retweets denoted by RT at the beginning of the tweet, removes the number characters, and removes punctuation marks like (.), (,), (:), (;), (?), and (!) ... The stages for cleaning text are divided into several steps including:

3. *Stop Word Removal*

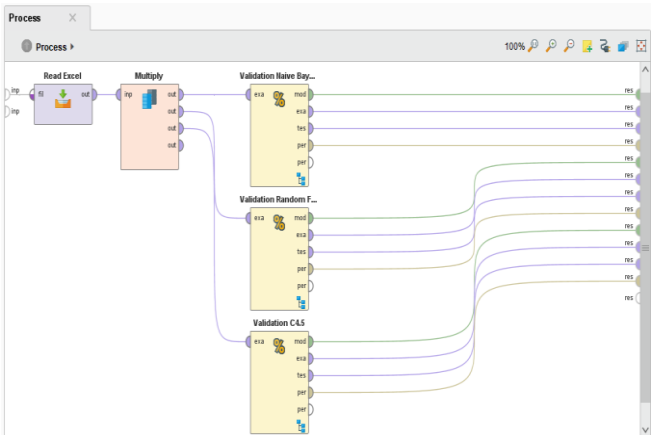
Based on the results of the previous tokenization, the next step is stopword removal, which is to eliminate stopword words or non-important words to be processed. At this stage, the application will select the word according to the word stopword in the stopword table. The word including the stop word used in this study is based on the stop word list taken fro https://github.com/masdevid/ID-Stopwords/blob/master/id.stopwords.02.01.2016.txt, where stop word refers to research (Tala, 2003). After passing the stopword removal process, the single word that successfully passes will be selected again with the sentiment table. This is done to facilitate the creation of frequency tables. Then the next stage will be weighted TF-IDF against existing data tweets. The final result of this stage is in the form of Bag of Words.

4. *Stemming*

At this stage, changing the token that is added into a basic word, by removing all of the affixes in the token. The list of words used in this study uses the dictionary stemming obtained from https://github.com/sastrawi/sastrawi/wiki/Stemming-Bahasa-Indonesia.

5. Modelling

The next stage after the stage of data preparation in CRISP-DM is modeling, based on the dataset obtained from the pre-processing process a process design will be used in this study. The following classification model in this study uses the naïve bayes algorithm as one of the classifiers in this study. The following model designs are applied in RapidMiner tools by designing the following process:

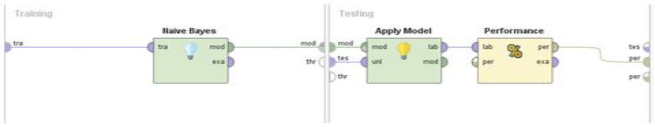


Source: Research result (2018)

Figure IV: Design of the Classification Model

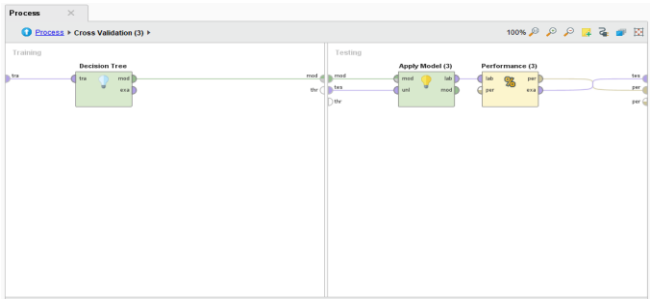
6. *Evaluation*

The next stage after the stage of data preparation in the CRISP-DM research model the evaluation, in this evaluation phase it aims to determine the usefulness value of the model that was successfully made in the previous step. For evaluation, 10-fold cross validation is used. The following process designs are used:



Source: Research result (2018)

Figure 5.
Design of the Validation Process for Naive Bayes



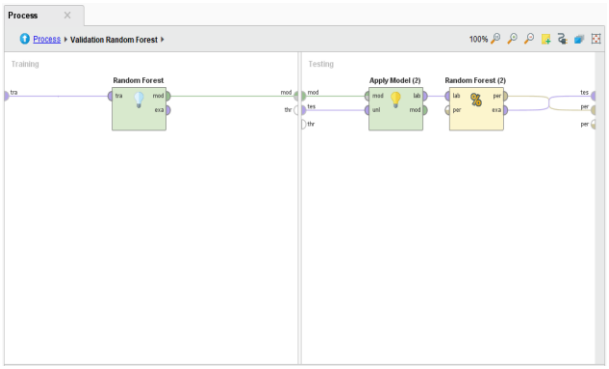
Source: Research result (2018)

Figure 6. Design of the Validation Process for C4.5

Validation operators will do this 10-fold validation process for each of the Naive Bayes and C4.5 algorithms. Each of the naïve bayes and k C4.5

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.235
e-ISSN : 2541-2019
p-ISSN : 2541-044X

experiments will be calculated for accuracy. Final accuracy is the average value of the accuracy of the ten trials. The results can be presented in the form of a confusion matrix and ROC curve.
The following is the experimental confusion matrix of each algorithm for each dataset.



Source: Research result (2018)

Figure 7.
Design the Validation Process for Random Florest

Validation operators will do this 10-fold validation process for each algorithm of naive bayes and Random Forest. Each of the naïve bayes and Random Forest experiments will be calculated for accuracy. Final accuracy is the average value of the accuracy of the ten trials. The results can be presented in the form of confusion matrix and ROC curve.
The following is the experimental confusion matrix of each algorithm for each dataset.

Curve Evaluation ROC (*Receiver Operating Characteristic*)

a. ROC Curve for the Naïve Bayes Algorithm on the dataset

Receiver Operating Characteristic (ROC) curves are another way to evaluate the accuracy of classification in a visual form (Curve). The following is the ROC curve for the Naïve Bayes algorithm on the dataset with an AUC accuracy value of 0.834 with a diagnosis level *Good Classification.*
ROC Curve for C4.5 Algorithm on the dataset
The following in the figure is the ROC curve for the Naïve Bayes algorithm on the @GrabID dataset with an AUC accuracy value of 0.500 with a diagnostic level Failure Classification

Source: Research result (2018)

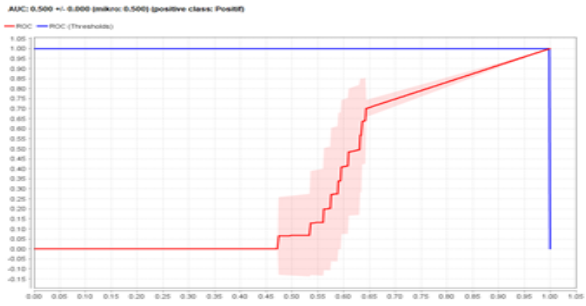
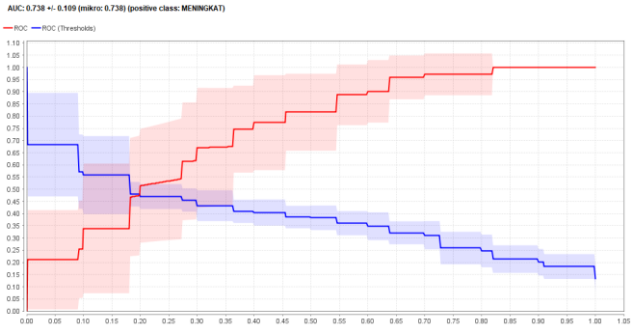


Figure 8.
The AUC Value in the ROC Graph C4.5 Algorithm

b. ROC Curve for Random Forest Algorithm on the dataset



Source: Research result (2018)

Figure IX.
AUC value in ROC Graph Random Florest algorithm

VII. CONCLUSION

In this study the authors classified the text to analyze sentiment from community tweets regarding online motorcycle taxi services. The research uses naïve bayes, c4.5 and random forest classification algorithms.

There are differences in the value of accuracy between the algorithms applied. In this study using datasets originating from the @GrabID account and @gojekindonesia. Accuracy produced in this study, resulted in a naïve bayes calcification algorithm having an accuracy value of 69.18% and AUC value of ¬0.771 so that included in fair classification, for random forest algorithm has an accuracy of 66.34% and AUC value of 0.738 so that it is included in fair clasification, while for the c4.5 algorithm has an accuracy of 65% and an AUC value of 0.686 so that it belongs to poor classification. From these result it can be concluded that

Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.235
e-ISSN : 2541-2019
p-ISSN : 2541-044X

the naïve bayes algorithm has higher accuracy compared to random forest and c4.5 algorithms, so that the difference in accuracy between naïve bayes and random forest is 2.84%, while the difference between naïve bayes and c4.5 is 3 , 53%. So that it can be concluded in this study the algorithm that has the best performance is the Naïve Bayes algorithm.

VIII. REFERENCES

Andriani, A. (2013). Sistem Pendukung Keputusan Berbasis Decision Tree Dalam Pemberian Beasiswa Studi Kasus : Amik " Bsi Yogyakarta ." *Seminar Nasional Teknologi Informasi Dan Komunikasi 2013 (SENTIKA 2013), 2013*(Sentika), 163–168.

Aprilla, D, Baskoro, Donny Aji, Ambarwati, Lia, & Wicaksana, IWayan Simri. (2013). Belajar Data Mining dengan Rapid Miner. Jakarta.

Dey, Lopamudra, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwari. (2016). *Sentiment Analysis of Review Datasets Using Naïve Bayes'and K-NN Classifier*. I.J. Information Engineering and Electronic Business. DOI: 10.5815/ijieeb.2016.04.07

Feldman, R dan Sanger, J., (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Pres : New York.

Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques*. Verlag Berlin Heidelrbeg: Springer

Hadi, Alfian Futuhul, Dimas Bagus C. W, Moh. Hasan. (2017). *Text Mining* Pada Media Sosial Twitter Studi Kasus: Masa Tenang Pilkada DKI 2017 Putaran 2. Seminar Nasional Matematika dan Aplikasinya, 21 Oktober 2017.Surabaya, Universitas Airlangga.

Liu, B. 2010. *Handbook of Natural Language Processing, chapter Sentiment Analysis and Analysis, 2nd Edition. Chapman & Hall / CRC Press.*

Negara, Edi Surya. Ria Andryani dan Prihambodo Hendro Saksono, 2016. *Twitter Data Analytics: Geospatial Data Extraction and Analysis, Pew Research Center*, INKOM, Vol. 10, No. 1, Mei 2016.

Nugroho, Didik Gabian, Yulison Herry Chrisnanto dan Agung Wahana. (2016). Analisis Sentimen Pada Jasa Ojek *Online* Menggunakan Metode Naïve Bayes. ISBN 978-602-99334-5-1. Prosiding SNST ke-7.

Prasetyo, Eko. (2014). Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab. Yogyakarta: Andi Offset.

Putler, Daniel S, dan Robert E Krider. (2015). *Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R*. CRC Press. USA. ISBN-13:978-1-4987-5970-0 (EPUB)

Sartika, Dewi dan Dana Indra Sensuse. (2017). Perbandingan Algoritma Kasifikasi Naïve Bayes dan Neares Neighbour dan Decission Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian. Jatisi. Vol 8, No. 2, Maret 2017.

Susanti, Aisah Rini, Taufik Djatna dan Wisnu Ananta Kusuma. (2017). *Twitter's Sentiment Analysis on Gsm Services using Multinomial Naïve Bayes*. Jurnal; TELKOMNIKA. Vol.15, No.3, September 2017. ISSN: 1693-6930, accredited "A" by DIKTI, Decree No: 58/DIKTI/Kep/2013.

Tala Fadillah Z. 2003. A study of stemming effects on Information retrieval in

Vercellis, C. (2009). *Business Intelligent: Data Mining and Optimization for Decision Making*. Southern Gate: John Willey & Sons Inc.