Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 1, October 2020
DOI : https://doi.org/10.33395/sinkron.v5i1.10565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Prediction of Netizen Tweets Using Random Forest, Decision Tree, Naïve Bayes, and Ensemble Algorithm

Yan Rianto[1)], Antonius Yadi Kuntoro[2)]
[1)2)]STMIK Nusa Mandiri, Jakarta, Indonesia
[1)]y-rianto@yahoo.com, yan.rianto@gmail.com, [2)]antonius.aio@nusamandiri.ac.id

Submitted : Jun 25, 2020 | **Accepted** : Sep 13, 2020 | **Published** : Oct 6, 2020

Abstract: The current Governor of DKI Jakarta, even though he has been elected since 2017 is always interesting to talk about or even comment on. Comments that appear come from the media directly or through social media. Twitter has become one of the social media that is often used as a media to comment on elected governors and can even become a trending topic on Twitter social media. Netizens who comment are also varied, some are always Tweeting criticism, some are commenting Positively, and some are only re-Tweeting. In this research, a prediction of whether active Netizens will tend to always lead to Positive or Negative comments will be carried out in this study. Model algorithms used are Decision Tree, Naïve Bayes, Random Forest, and also Ensemble. Twitter data that is processed must go through preprocessing first before proceeding using Rapidminer. In trials using Rapidminer conducted in four trials by dividing into two parts, namely testing data and training data. Comparisons made are 10% testing data: 90% Training data, then 20% testing data: 80% training data, then 30% testing data: 70% training data, and the last is 35% testing data: 65% training data. The average Accuracy for the Decision Tree algorithm is 93.15%, while for the Naïve Bayes algorithm the Accuracy is 91.55%, then for the Random Forest algorithm is 93.41, and the last is the Ensemble algorithm with an Accuracy of 93, 42%. here.

Keywords: Decision Tree, Naïve Bayes, Random Forest, Set, Twitter

INTRODUCTION

Twitter is an online social networking and microblogging service that allows users to send and read text-based messages (Wikipedia, 2019). As well as with the official Twitter account of the DKI Jakarta governor, @aniesbaswedan. Mr. Anies Rasyid Baswedan and Sandiaga Salahuddin Uno were the Governor and Deputy Governor of DKI Jakarta Province for the period 2017-2022. The number of news about the governor including Tweets posted on the account of Mr. Anies Rasyid Baswedan, both those that have Positive, Negative, and Neutral, as in Figure 1 and Figure 2 below:

Source: Electronic media
Fig. 1 Example of news about Governor DKI Jakarta

*name of corresponding author

Source: @anisbaswedan Twitter account
Fig. 2 Examples of Tweets from Netizens

This study will retrieve data or Tweets posted by Netizens on Mr. Anies Rasyid Baswedan's official Twitter account, @aniesbaswedan. Retrieval of data using the Rapidminer application, carried out effectively and efficiently which is then carried out labeling Positive and Negative by third parties, namely by 100 respondents. Labeling is useful for analyzing one's opinion, one's evaluation, one's attitude, and one's emotions into written language, in this case, it can be called sentiment. One of the disciplines that studies methods for extracting knowledge or finding patterns from large data are data mining. Data mining is the process of extracting to obtain important information that is implicit and previously unknown, from data (Witten et al., 2011). Huge of interesting (non-trivial, implicit, previously unknown, and potentially useful) patterns or knowledge from huge amounts of data (Jiawei Han & Kamber, 2013). Data mining is often considered as part of Knowledge Discovery in Database (KDD), which is a process of finding useful knowledge from data. Besides, data mining is also known as knowledge extraction, pattern analysis, information harvesting, and Business intelligence.

There are 5 main roles of data mining, namely: Estimation, Prediction, Classification, Classification, and Association. Data mining algorithms that are often used in classification include Naïve Bayes, K-Nearest Neighbors, Decision Trees, ID3, CART, Linear Discriminant Analysis, Logistic Regression, Ensemble, and others. However, in this study, the author will only use the Random Forest, Decision Tree, Naïve Bayes, and Ensemble algorithms to process, classify, and my knowledge from the Twitter dataset on the @aniesbaswedan account.

In data mining, research on the classification of Twitter posts has been conducted by other researchers. Most of these studies have focused on identifying predictor variables. There is a lot of research in previous literature that explains what factors can maximize the classification process of Twitter's post data. These factors are generally divided into two, namely Twitter data processing pre-processing factors and Twitter data mining.

Data posting on Twitter or so-called Tweet (tweet) is very susceptible to noisy data, missing or incomplete data, and inconsistent data because usually, the Tweet data for each post will be in different forms of writing and very heterogeneous. For this reason, good, adequate, and representative data must be prepared as a first step that cannot be ignored. The reliability of the information that will be extracted from an existing database depends on the quality of the data that will be processed. There are several data preprocessing techniques that can be used to produce quality data. Data cleaning can be applied to eliminate noise and inconsistent data. Data reduction can reduce data size. Data transformation to improve the accuracy and efficiency of data mining algorithms that involve distance measurement (Jiawei Han & Kamber, 2013).

Many studies that make regional leaders or leaders of the country and also public figures become the object of research and most of them discuss sentiment analysis or sentiment analysis, using the theory of text mining. While in this study, the author tries to conduct research using Twitter data on the official Twitter account of @aniesbaswedan to predict whether Netizens tend to Tweet Positive or Negative news. In this study will be compared to the algorithms of Random Forest, Decision Tree, Naïve Bayes, and Ensemble to get the best accuracy from the predictions above.

The T-test of this research is to get assumptions against the Twitter user ID whether to lead to the Positive or Negative sentiment of the opinion posted on Twitter's official account @aniesbaswedan, using Random Forest algorithms, Decision Tree, Naïve Bayes, and Ensemble.

The scope of this research is to Cultivate, perform a comparison of algorithms classification algorithms of Random Forest, Decision Tree, Naïve Bayes, and Ensemble. The data used in this study is public comment data citizens on official Twitter @aniesbaswedan.

Based on the above identification of the problem, this research seeks to answer which model question is more accurate and precise between the Random Forest algorithm, Decision Tree, Naïve Bayes, and Ensemble in the prediction of user ID or netizen that tweets Positive and Negative terms research.

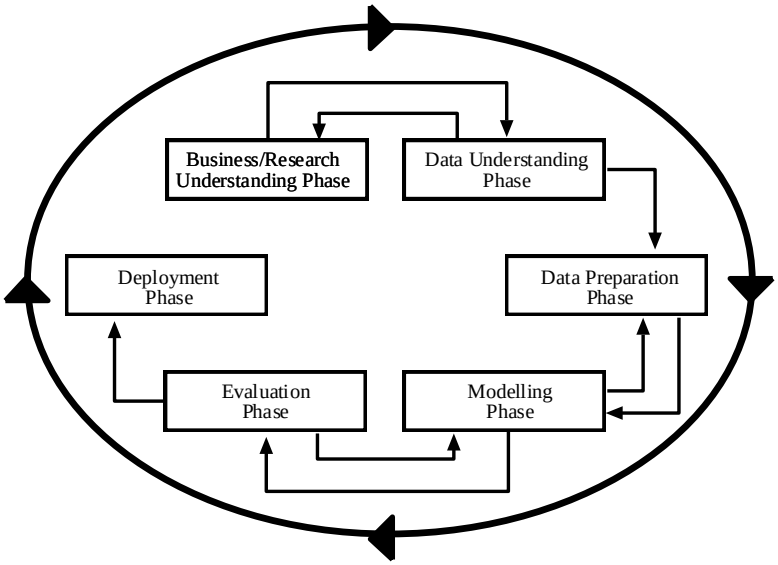
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 1, October 2020
DOI : 10565https://doi.org/10.33395/sinkron.v5i1.10565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

LITERATURE REVIEW

Data mining

Data mining is the process of discovering interesting patterns and knowledge of large amounts of data (Jiawei Han & Kamber, 2013). Data Mining can be seen as a science that explores datasets in large quantities for the excavation of implied information, previously unknown, and potentially generating useful information (Gorunescu, F., 2011). According to (Larose, D. T., 2005) Data Mining is a process of discovering meaningful, ber-patterned, and tendency relationships by examining in a large set of data stored in storage using pattern-recognition techniques such as statistical and mathematical techniques.

The process of Data Mining must comply with the process procedure CRISP-DM (Cross-Industry Standard process for data mining) namely as the whole process, preprocessing data, forming models, evaluation model, and finally the deployment of models (Larose, D. T., 2005). CRISP-DM provides the standard data mining process as a general troubleshooting strategy of the Business or research unit. Can be seen in Figure 3 as follows:



Source: Images of CRISP-DM process (Larose, 2005)
Fig. 3 CRISP-DM process

Algorithm

1. The Random Forest algorithm is a result of the development of the Decision Tree, where each Decision Tree has been conducted training using individual samples and each attribute is broken down on tree a selected Tree between a random subset.
2. The Decision Tree algorithm is an algorithm in the Decision Tree method that converts data into decision trees by using the entropy calculation formula.
3. Naïve Bayes ' algorithm is one of the methods of machine learning that utilizes probability calculations by predicting future probability based on previous experience.
4. The Ensemble algorithm uses several learning algorithms to achieve better prediction solutions than algorithms that can be obtained from one of the constituent learning algorithms and usually allows to become even more flexible structures available among the models themselves. The voting feature will make predictions with the maximum sound/recommendation of various predictive models while predicting the outcome of a classification problem.

Twitter dan Netizen

1. Twitter is a popular microblogging service, users can post status or messages referred to as tweets of no more than 140 characters. Most of the cases, users write down their messages much less than the specified character constraints. Twitter is one of the social media's largest and most dynamic data contributors based on user-generated content. It is estimated that around 200 million Twitter users post 400 million tweets per day. Tweets This can be an expression of opinion on a wide variety of topics that can help to share opinions on users, identification of irregularities, incidents that cause discomfort, prediction of political behavior, and sport, acceptance or rejection of politics and all communication expressed by word of Mouth (Da Silva, et al. 2014).
2. Netizen is a combination of the word internet and citizen that means the Internet or personal citizens who use the Internet in general and active in social media.

e-ISSN : 2541-2019
p-ISSN : 2541-044X

The Pareto Principle

The Pareto principle is also known as Rule 80-20, stating that for many occurrences, about 80% of its effect is caused by 20% of the cause. This principle is referred to by business management thinker Joseph M. Juran, who named him by Italian economist Vilfredo Pareto (Flux & Pareto, 1897). With this Pareto approach, authors will use it in the preprocessing process on the Data Set in Excel to define some new attributes that contain Active or Passive. This Pareto will be used in pre-preprocessing to obtain the active and passive netizen criteria tweeting on the account above. This calculation is used on a worksheet using MS Excel.

K-Fold Cross Validation Test

K-Fold Cross-validation is a validation technique by randomly dividing the data into k sections and each part will be done classification process (Han, J., & Kamber, M., 2007). Cross-validation is a statistical method used to evaluate and compare algorithms by dividing data into two segments, the first segment used as training data and the second segment as data testing in model validation (Witten et al., 2011). Data Training is the data that will be used in conducting learning while data testing is data that has never been used as a learning and will serve as a data of truth testing or accuracy of learning outcomes (Witten et al., 2011).

Using the K-fold Cross-validation, also known as the 10-fold cross-validation test will be conducted as much as K. Results of extensive experiments and theoretical evidence, indicating that the 10-Fold Cross-validation is the best choice for obtaining accurate validation results. Therefore, in the general test, the value of K is done 10 times. 10-Fold Cross-Validation will repeat the test 10 times and the measurement result is the average value of 10 test times.

Study Review

1. Research conducted by (Cureg et al., 2019) titled "Sentiment Analysis on Tweets with Punctuations, Emoticons, and negations". In this study, tweets that included different parameters (emoticons, negations, and punctuation) were gathered and explained by experts to label their sentiments. Machine learning was applied in this study to formulate an optimal model. The test results show that the included features provide significant performance to identify the sentiment of a given microblog statement. The algorithms that are used to build models are KNN and naïve Bayes with English and Filipino language data. With the results of the research on the Calculating Kappa is K-NN algorithm with 40% accuracy, **32.5% Naïve Bayes** algorithm, and SVM algorithm is 38.77%.
2. Research conducted by (Kartiko & Sfenrianto, 2019) entitled "Accuracy for Sentiment Analysis of Twitter Students on E-Learning in Indonesia using Naïve Bayes Algorithm Based on Particle swarm optimization". This research aims to analyze the accuracy of students ' sentiments of E-Learning that use Bahasa Indonesia in social media Twitter both Positive and Negative opinions. The algorithm used is **naïve Bayes (NB)**. Then to optimize the accuracy of the calculation result, the Particle swarm Optimization (NB-PSO) approach is used. To optimize accurate results, this study uses three experimental sequences (scenario 1, Scenario 2, and Scenario 2) for NB and NB-PSO algorithms. Each scenario uses a different Positive and Negative comment. The results of the experiment showed that in scenario 1 increased accuracy was 10.00% for NB-PSO. Scenario 2 There is a 13.33% increase in accuracy on NB-PSO. Meanwhile, in scenario 3 the improved accuracy is 27.22% for NB-PSO. This result proves that the accuracy of NB-PSO is better than NB for all scenarios.
3. Research conducted by (Al-Rubaiee et Al., 2016) entitled "Analysis of the Relationship Between Saudi Twitter Posts and the Saudi Stock Market". In this paper, Twitter has been selected as a platform for mining opinion in trading strategies with the Saudi stock market to implement and illustrate the relationship between the Saudi tweets (i.e. standard dialects and the Arabian Gulf) and the Saudi market index. Our knowledge, this is the first study conducted in the Saudi Tweet and the Saudi stock market. With accuracy on the algorithm, **Naïve Bayes** of **69.86%** AND THE accuracy of the SVM algorithm is 96.60%
4. The research was done by (Blatnik et al., 2014) titled "Movie sentiment analysis based on public Tweets". In the study, using the Python programming language with the NLTK library and comparing the results obtained by traditional machine learning techniques using rapid miner. Our focus is on Twitter -a microblogging platform with a maximum of 140 characters per post (Tweet), more specifically, on the collection of sentiments for a particular film. In this research, the results achieved have timely accuracy of 93% and late accuracy of 71%. With algorithm accuracy results **74.14% Naïve Bayes** and the k-NN algorithm is 77.59%.
5. Research conducted by (Buntoro, 2017) titled "The sentiment analysis of the prospective governor of DKI Jakarta 2017 on Twitter". This research is expected to be useful to help to research public opinion that contains Positive sentiment, Neutral, or Negative. The methods used in this study, for preprocessing data using tokenization, cleansing, and filtering, to define sentiment classes with Lexicon Based methods. For its classification process using the Naïve Bayes classifier (NBC) method and Support Vector Machine (SVM).

The Data used is a Tweet in Bahasa Indonesia with the keyword Ahy, Ahok, Anies, with the number of datasets as much as 300 tweets. The result of this research is the analysis of the sentiment on the prospective governor of DKI Jakarta 2017. Highest accuracy is gained when using the **Naïve Bayes** classifier (NBC) classification method, with an average value of accuracy reaching **95%**, the precision value is 95%, THE value of recall 95% value OF TP rate of 96.8% and a TN rate value of 84.6%.

6. Heart disease predictions using Ensemble - Weighted Vote - based Machine Learning methods written by (Alhamad et al., 2019) states that the implementation of Machine Learning methods in the public Data Set (Cleveland, Hungary, Switzerland, VA Long Beach, & Stat log) is commonly used by researchers for heart disease predictions, including the development of its objecting tools, still not handling missing value, noisy data, unbalanced class, and even data validation efficiently. Therefore, the mean/mode approach is proposed to handle missing value replacement, Min-Max normalization to handle Smoothing Noisy data, K-Fold Cross Validation to handle data Validation, and the Ensemble approach uses a Weighted Vote (WV) method that can unify the performance of each Machine learning method to take classification decisions at once to reduce unbalanced classes. The results showed that the proposed method gave an accuracy of 85.21%, to improve the accuracy of the machine learning methods, the difference of 7.14% with Artificial Neural Network, 2.77% with Support Vector machine, 0.34% with Decision Tree, 2.94% with Naïve Bayes, and 3.95% with K-Nearest Neighbor.
7. Similarly, the acquisition of the 2013 curriculum sentiment analysis on Twitter uses the Ensemble Feature and the K-Nearest Neighbor method, by (Mentari et al., 2018). In the study, sentiment analysis was conducted to find out which of these emerging opinions were divided into Positive opinions or Negative opinions. The features and methods used are the Ensemble feature and the K-Nearest Neighbor (K-NN) Classification method. Ensemble feature is a combined feature, a statistical feature of Bag of Words (BoW) and semantic (Twitter specific, textual features, PoS features, Lexicon based features). Based on a series of tests, the combination features impacting method accuracy K-Nearest Neighbor (K-NN) to determine the opinion Positive or other Negative. Combining this feature can complement the disadvantages of each feature so that the final result of accuracy obtained by combining both features is **96%**. Different things if only use the feature independently only, the accuracy obtained only reached 80% in the feature Bag of Words (bow) and 82% in Ensemble features without Bag of words (bow).
8. The study entitled A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining by (Ratul & Engineering, nd) in 2020 is to ensure general security including prevention of city crime by applying several classification algorithms such as Random Forests, Decision Trees, There are Boost, Extra Tree Classification, and K-Neighbors Classification, and 4 Ensemble Models. The results obtained. Random Forest, Decision Tree, and Ensemble Model 1, 3, and 4 algorithms produce 90% accuracy.
9. A study titled Comparison of Performance of Various Data Classification Algorithms with Ensemble Methods Using Rapidminer by (Puyalnithi et al., 2016), which examines the impact of various classification algorithms in predictions of unknown label attributes. The predictions used in the study are naïve Bayes, Decision Tree, and Random Forest using Rapidminer. The results of the accuracy obtained are **naïve Bayes 84.34%, Random Forest 89.96%, and Decision Tree 89.97%.**
10. Another study, conducted in 2019 under the title Sentiment Analysis of the Indonesian Police Mobile Brigade Corps Based on Twitter Posts Using the SVM And NB Methods by (Pratama et al., 2019) describes the superior SVM algorithm with 86.96% accuracy while **naïve Bayes** algorithm with **86.48%**accuracy. The value of this accuracy in the can of research with the data object is the Brimob Corps on the social media Twitter to be analyzed whether tweets posted are Positive or Negative.

METHOD

Business / Research Understanding Phase

This stage is an understanding of the research object. In this research, the author uses Twitter data from the official account of the governor of DKI Jakarta, namely @aniesbaswedan in the period 28 Sept 2019 S/d 09 November 2019. Twitter data Retrieval Twitter using the app on Rapidminer. In this stage also done an understanding to find Positive and Negative labels on text posted by the user. In addition to the text label, it can also be obtained Active and Passive from Twitter users.

Data Understanding Phase

This stage is the process of understanding the data that will be used as material to be researched to be done to the following stage namely Preprocessing. Below are the steps that will be done.

Set up a total Data Set from personal account Twitter @aniesbaswedan and downloaded data with 29,340 tweets, tweet data in download using the tools from Rapidminer, then continue in Excel format. Datasets already stored in Excel are further processed to identify duplicate tweets in the post, in other words, the Dataset is cleaned with a process called data cleaning. After Cleaning is obtained as much data 12,027 can be used. In this study only took 10,000 data consisting of a Positive label of Positive 5,000 and Negative label as much as 5,000 data. This

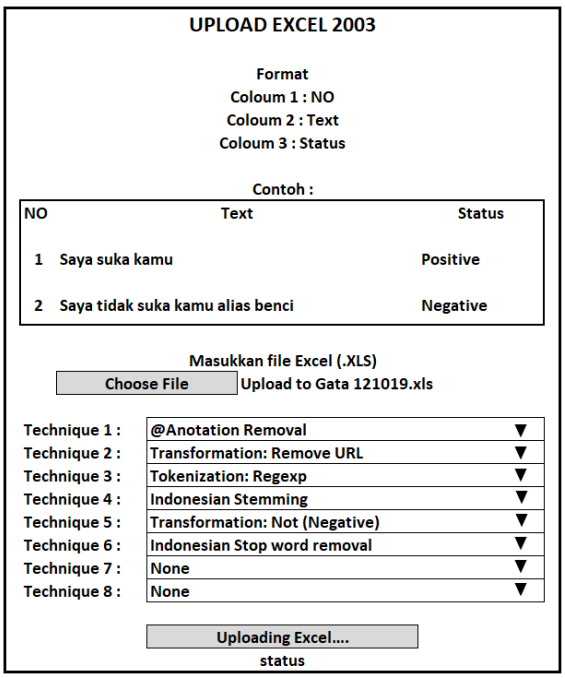
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 1, October 2020
DOI : 10565https://doi.org/10.33395/sinkron.v5i1.10565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

labeling involved 100 respondents using the Crowdsourced labeling method, which is the data labeling method involving the participation of the general audience. The process of labeling for datasets that do not require special skills or to study participants in giving labeling (Rachmat & Lukito, 2016). Many respondents will accelerate the giving labeling process and will also be more neutral in the label. Another thing that also benefits writers from the many respondents in this process, is that there is no need for a large fee when compared to using the Help of experts to do so.

Data Preparation and Modelling

The next step is to prepare the data before the data will be modeling or called Data Preparation. For this 2nd stage of preparing the data to perform the steps known as text preprocessing, using two preprocessing applications, first using Gataframework accessed via Link http://Gataframework.com/textmining that can be used for free is also easy to use because it does not have to create an account to use the foods and continued preprocessing of Rapidminer.

The next stage is the preprocessing of Rapidminer in the order shown in Figure 4 below:



Source: Gataframework Tools

Fig. 4 Gataframework display Images

1. @Anotation Removal

   The first step of this is the text decomposed by white space, all the annotations contained in the Tweet will be eliminated and the lower case or change the letter in the text to all lower case, such as the following sample Table 1:

Table 1
Table Comparison of Text before and after @Anonation Removal Process

Text	@Anotation Removal
Tenangkan Massa @ganjarpranowo Turun ke Tengah Mahasiswa, kalau @aniesbaswedan? #AniesGaBener https://t.co/t9lqYpmGLx	tenangkan massa turun ke tengah mahasiswa, kalau #aniesgabener https://t.co/t9lqypmglx

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 1, October 2020
DOI : 10565https://doi.org/10.33395/sinkron.v5i1.10565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

@aniesbaswedan @DKIJakarta @pln_123 @PT_TransJakarta @DishubDKI_JKT @dinaslhdki Hati hati Pak @aniesbaswedan sampai saat ini saja wakilnya belum ada... kasihan @PKSejahtera di zholimi terus. https://t.co/K294APLorx	hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus. https://t.co/k294aplorx
@wong_sedeng @prahar_77 @PSI_Jakarta @rianernesto @psi_id @aniesbaswedan Siapapun presiden Indonesia... Pasti Ngutang .. gubernur nya aja ngutang ?????? https://t.co/wphXQRhFMw	siapapun presiden indonesia... pasti ngutang .. gubernur nya aja ngutang ?????? https://t.co/wphxqrhfmw
@sabar_mbok @IndoPluralitas @alvaro3_lee3_ @aniesbaswedan @DKIJakarta Alhamdulillah, mudah2an dosa2nya Pak Anies terhapus karena fitnah2 ini. https://t.co/Wk1KtLkTvJ	alhamdulillah, mudah2an dosa2nya pak anies terhapus karena fitnah2 ini. https://t.co/wk1ktlktvj

2. Transformation: Remove URL
Often a URL appears from Twitter data Twitter making data ineffective and meaningless. For that it is necessary to delete URL the URL or bias also to remove Internet links, such as the following Table 2:

Table 2
Table Comparative Text before and after the @ Transformation process: Remove URL

Text	**Transformation: Remove URL**
tenangkan massa turun ke tengah mahasiswa, kalau #aniesgabener https://t.co/t9lqypmglx	tenangkan massa turun ke tengah mahasiswa, kalau #aniesgabener
hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus. https://t.co/k294aplorx	hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus.
siapapun presiden indonesia... pasti ngutang .. gubernur nya aja ngutang ?????? https://t.co/wphxqrhfmw	siapapun presiden indonesia... pasti ngutang .. gubernur nya aja ngutang ??????
alhamdulillah, mudah2an dosa2nya pak anies terhapus karena fitnah2 ini. https://t.co/wk1ktlktvj	alhamdulillah, mudah2an dosa2nya pak anies terhapus karena fitnah2 ini.

3. Tokenization: Regexp
The tokenization process is performed after the transform cases. All unnecessary characters will be discarded. Includes excessive white space and all punctuation. This process will be done on any documents entered from the document collection. So it is obtained a unique word and can represent documents, such as the following Table 3 example:

Table 3
Table Comparison of Text before and after the @ tokenization process: regexp

Text	**Tokenization: Regexp**
tenangkan massa turun ke tengah mahasiswa, kalau #aniesgabener	tenangkan massa turun ke tengah mahasiswa kalau aniesgabener
hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus.	hati hati pak sampai saat ini saja wakilnya belum ada kasihan di zholimi terus

siapapun presiden indonesia... pasti ngutang .. gubernur nya aja ngutang ??????	siapapun presiden indonesia pasti ngutang gubernur nya aja ngutang
alhamdulillah, mudah2an dosa2nya pak anies terhapus karena fitnah2 ini.	alhamdulillah mudahan dosanya pak anies terhapus karena fitnah ini

4. Indonesian Stemming
After the result of the transformation not Negative will be followed by the steaming process is to remove the suffix that is found in each word so that it is a basic word using Indonesian stemming for a Tweet - speaking Indonesia, such for example Table 4 follows:

Tabel 4
Table Comparison of Text before and after the Indonesian stemming process

Text	Indonesian Stemming
tenangkan massa turun ke tengah mahasiswa kalau aniesgabener	tenang massa turun ke tengah mahasiswa kalau aniesgabener
hati hati pak sampai saat ini saja wakilnya belum ada kasihan di zholimi terus	hati hati pak sampai saat ini saja wakil belum ada kasihan di zholimi terus
siapapun presiden indonesia pasti ngutang gubernur nya aja ngutang	siapa presiden indonesia pasti ngutang gubernur nya aja ngutang
alhamdulillah mudahan dosanya pak anies terhapus karena fitnah ini	alhamdulillah mudah dosa pak anies hapus karena fitnah ini

5. Transformation: Not (Negative)
From the results of Tokenization (Regexp), the next process is transformation not Negative. For this example, in the text used previously, there was no change because there were no words made by Transformation Not Negative. But to clarify the purpose of the process, another text from the same local data is used, such as the example in Table 5 below:

Tabel 5
Table Comparison Of Text Before And After The Transformation Process: Not (Negative)

Text	Transformation: Not (Negative)
untuk erat tali saudara yg cerai berai oleh radikalisme pak gimana pak jawab saya udah keren belum	untuk erat tali saudara yg cerai berai oleh radikalisme pak gimana pak jawab saya udah keren belum_
tanya saya simple anda boleh orang dagang trotoar yang notabene buat untuk jalan kaki bukan untuk dagang iya atau tidak	tanya saya simple anda boleh orang dagang trotoar yang notabene buat untuk jalan kaki bukan_untuk dagang iya atau tidak_
lo me bicara harga juga semua dapat harga lo liat lapang sini lo orang jakarta bukan	lo me bicara harga juga semua dapat harga lo liat lapang sini lo orang jakarta bukan_
lha itu jpo sdh ada atap dul ngapain di lepas lain halnya kalo emang dr dolo ga ada kita trima dgn lapang dada buat bijak itu yg manfaat jgn buat bijak yg lebih tdk manfaat dr belum	lha itu jpo sdh ada atap dul ngapain di lepas lain halnya kalo emang dr dolo ga ada kita trima dgn lapang dada buat bijak itu yg manfaat jgn buat bijak yg lebih tdk manfaat dr belum_

6. Indonesian Stop Words removal
This stop word stage will refine the token by length filter Stage. Words consisting of more than 3 letters and included in the stop words will be discarded. Because the word does not reflect the contents of the document even though it frequently appears, such as Table 6 example follows:

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 1, October 2020
DOI : 10565https://doi.org/10.33395/sinkron.v5i1.10565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Tabel 6
Comparative Text before and after the Indonesian Stop Word removal process

Text	Indonesian Stop word removal
tenang massa turun ke tengah mahasiswa kalau aniesgabener	tenang massa turun mahasiswa aniesgabener
hati hati pak sampai saat ini saja wakil belum_ada kasihan di zholimi terus	hati hati wakil belum_ada kasihan zholimi
siapa presiden indonesia pasti ngutang gubernur nya aja ngutang	presiden indonesia ngutang gubernur ngutang
gw maklum lah lho pantas aja lho mati nyinyir pak sahabtanya dan buzzer lain	maklum lah lho mati nyinyir sahabtanya buzzer

RESULT

Model Decision Tree experiments and testing results

Of the 10,000 Text data that was posted and processed using the Decision Tree algorithm on Rapidminer with data testing comparison and data training 10:90 There are as many as 4152 data in positive predictions and fact positive, 4128 negative predicted data and reality negative, 372 data predicted positive but Negative and 348 negative predicted data But reality Positive, as in the following Table 7 below:

Table 7
Confusion Matrix Decision Tree data testing 10% and data training 90%.
accuracy: 92.00%

	true Positive	true Negative	class precision
pred. Positive	4152	372	91.78%
pred. Negative	348	4128	92.23%
class recall	72.72%	91.73%	

The ROC curve measurement by using the UnderCurve Area (AUC) resulting in an AUC value of 0. 957, as in Figure 5 below.

Source: Rapidminer Tools
Fig. 5 Images Under Curve Area graph (AUC) of Decision Tree, data testing 10% and training Data 90%

Test results of Naïve Bayes Model experiments and testing

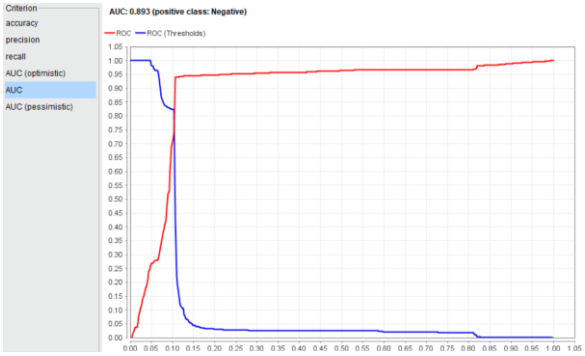
Accuracy value obtained by testing 10% Data Testing comparison: The training data is 90%; Accuracy = 91.64%. Of the total 10,000 datasets were processed, as many as 4024 the amount of data predicted positive and positive, 4224 negative predicted data and negative, 276 predicted data positive but negative, and 476 negative predicted data But Positive as in Table 8 below.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 1, October 2020
DOI : 10565https://doi.org/10.33395/sinkron.v5i1.10565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 8
Confusion Matrix Naïve Bayes data testing 10% and data training 90%
accuracy: 91.64%

	true Positive	true Negative	class precision
pred. Positive	4024	276	93.58%
pred. Negative	476	4224	89.87%
class recall	89.42%	93.87%	

ROC curve measurements using the Area Under Curve (AUC) which produces an AUC value of 0.893 as shown in Figure 6.



Source: Rapidminer Tools
Fig. 6 Images Under Curve Area graph (AUC) Naïve Bayes algorithm, 10% data testing, and 90% training Data.

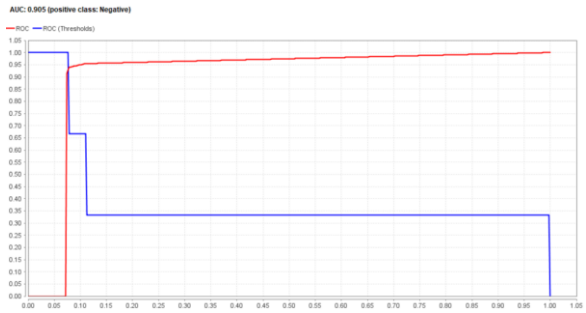
Results of experimental and Model Ensemble Vote

With a comparison of the 10% Data Set tester and the 90% training Data Set, it is generated by accuracy. Accuracy results of 91.44%. Of the total 10,000 datasets were processed, as many as 4152 the amount of data predicted positive and positive, 4224 negative predicted data and negative, 227 predicted data positive but negative, and 348 negative predicted data But Positive as in the Table 9 below:

Table 9
Confusion Matrix Ensemble, testing 10% data testing and 90% training Data
accuracy: 91.44%

	true Positive	true Negative	class precision
pred. Positive	4152	277	93.75%
pred. Negative	348	4224	92.39%
class recall	92.27%	93.84%	

ROC curve measurements using the Area Under Curve (AUC) which produces an AUC value of 0.905 as shown in Figure 7



Source: Rapidminer Tools
Fig. 7 Images Under Curve Area graph (AUC) of Ensemble algorithm, 10% data testing and 90% training Data

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 1, October 2020
DOI : 10565https://doi.org/10.33395/sinkron.v5i1.10565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

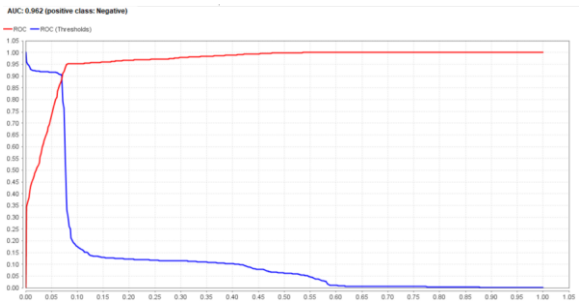
The results of the experiment and test Model of Random Forest

With a comparison of the 10% Data Set tester and the 90% training Data Set, it is generated by accuracy. Accuracy results of 93.08%. Of the total 10,000 datasets are processed, as many as 4153 amount of data predicted positive and positive, 4224 negative predicted data and negative, 226 predicted data positive but negative, and 347 negative predicted data But reality positive as in Table 10 below:

Table 10
Confusion Matrix Random Forest, testing 10% data testing and 90% training Data
accuracy: 93.08%

	true Positive	true Negative	class precision
pred. Positive	4153	276	93.77%
pred. Negative	347	4224	92.41%
class recall	92.29%	93.87%	

ROC curve measurements using the Area Under Curve (AUC) which produces an AUC value of 0.962 as shown in Figure 8.



Source: Rapidminer Tools

Fig. 8 Images Under Curve Area graph (AUC) of Ensemble algorithm, 10% data testing and 90% training Data

DISCUSSIONS

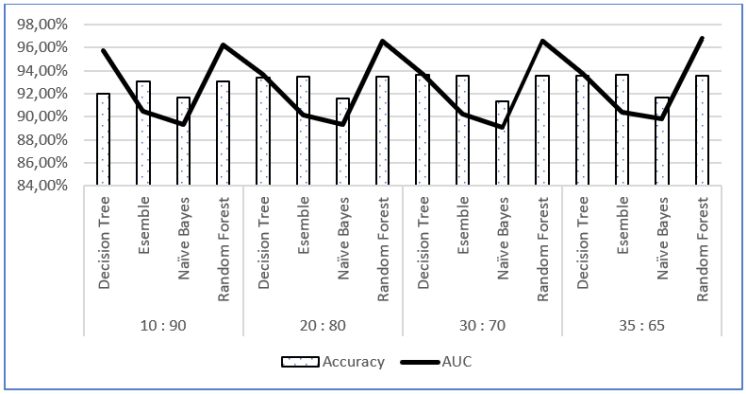
Overall Model Comparison

The results of processing with Rapidminer above as a representative of 4 comparisons of data testing and data training in each algorithm. From table 3.7, we can see the comparison results of the four algorithms used in the research of this Random Forest, Decision Tree, Naïve Bayes, and Ensemble with Vote features, based on data sharing testing: The following data training; 10:90, 20:80, 30:70 and 35:65 on testing 10% data testing comparison and 90% training, the accuracy value of the Random Forest algorithm is 93.08% and higher than the other two algorithm, with the AUC value of 0.962. In comparative data testing 20% and training 80% training data, the accuracy value of the Random Forest algorithm is 93.45% with its AUC value of 0.966. In the last comparison with testing 30% Data testing and training of 70%, the Decision Tree algorithm with an accuracy of 93.60% and with an AUC value of 0937, the latter is a ratio of 35% to data testing and 65% for training data with the highest accuracy result in the Ensemble of 93.60% and the AUC of 0.904. For average overall experiments can be seen in Table 11 and Figure 9 below.

Table 11
Recapitulation for Rapidminer Test Data Set

Algorithm	Testing	Training	Accuracy	AUC
Decision Tree	10%	90%	92,00%	0,957
Together	10%	90%	93,06%	0,905
Naïve Bayes	10%	90%	91,64%	0,893
Random Forest	10%	90%	93,08%	0,962
Decision Tree	20%	80%	93,40%	0,937

Together	20%	80%	93,44%	0,901
Naïve Bayes	20%	80%	91,60%	0,893
Random Forest	20%	80%	93,45%	0,966
Decision Tree	30%	70%	93,60%	0,937
Together	30%	70%	93,57%	0,902
Naïve Bayes	30%	70%	91,30%	0,891
Random Forest	30%	70%	93,57%	0,966
Decision Tree	35%	65%	93,58%	0,938
Together	35%	65%	93,60%	0,904
Naïve Bayes	35%	65%	91,66%	0,898
Random Forest	35%	65%	93,55%	0,968



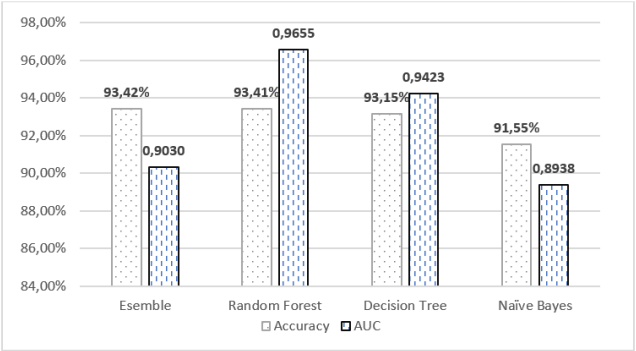
Source: Ms Excel

Fig. 9 Images Comparison of Rapidminer Experiment Results

Whereas if averaged from to four experiments on 4 algorithms, as in Table 12 and figure 10 follows.

Table 12
Results on average dataset Rapidminer test datasets.

Algorithm	Accuracy	AUC
Together	93,42%	0,903
Random Forest	93,41%	0,9655
Decision Tree	93,15%	0,9423
Naïve Bayes	91,55%	0,8938



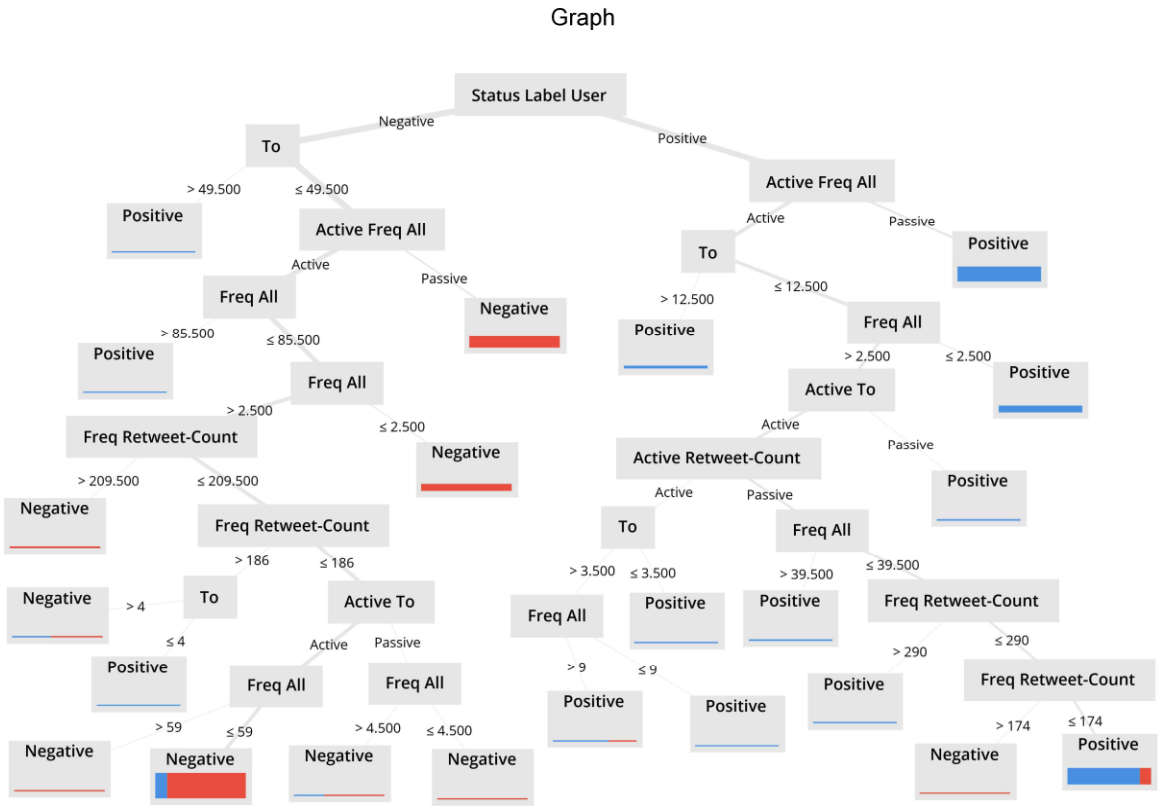
Source: Ms Excel

Fig. 10 Average comparison of Rapidminer Experiment Results

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 1, October 2020
DOI : 10565https://doi.org/10.33395/sinkron.v5i1.10565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In table 3.8, we can see that the average of the four experimental data obtained an average accuracy and the average Under Curve Area (AUC) of the Ensemble algorithm with an accuracy of 93.42% and AUC of 0.9030. For the Random Forest Algorithm with an accuracy of 93.41 and an AUC of 0.9655. For the Decision Tree Algorithm with 93.15% accuracy and AUC of 0.9423. As for the Naïve Bayes algorithm with an accuracy of 91.55% and AUC of 0.8938.

Tweet Prediction Analysis Netizen



Source: Rapidminer

Fig. 11 Image of Tree Diagram of a Random Forest model, with DataSet 2000

From the tree diagram in figure 11 above that can be from the Rapidminer process with a Random Forest model and the number of datasets as much as 2000 data. The Label Status of the User who is a netizen identity that posts a tweet can be predicted whether the tweet is Negative or Positive with the conditions and conditions as in the above image.

CONCLUSION

From the comparison of Random Forest algorithms, Decision Tree, Naïve Bayes, and Ensemble, from the trial with the distribution of data testing: Data training 10%: 90%, 20%: 80%, 30%: 70% and 35%: 65%. Algorithms Random Forest excelled on test tests with compositions Data testing 10%: Training data 90% with 93.08% accuracy and AUC of 0.962, then superior to Data testing 20%: Training data 80% with accuracy 93.45 and AUC 0.966. While in test data testing 30%: Training Data 70%, superior to the Decision Tree with an accuracy of 93.60% and AUC 0.937. The final test is on data testing 35%: Training data training 65%, obtained an accuracy of 93.60%, and AUC 0.904 for Ensemble algorithm.

The Random Forest algorithm can be predicted to affect the Twitter or netizen user ID whether The Tweet posted to the account @aniesbaswedan the majority leads to sentiment Positive or Negative. The things that affect your user ID or netizen lead to positive or negative such as netizen Tweet frequency, then the contents of a tweet are positive or negative, also the type whether tweet or re-tweet.

On existing data sets, accuracy is influenced by algorithm selection and data testing comparisons with training data. This is evidenced by 4 experiments carried out a result of 3 of 4 algorithms that can excel its accuracy. Other things that also take the rise of accuracy and AUC are pre-processing over downloaded data sets.

e-ISSN : 2541-2019
p-ISSN : 2541-044X

REFERENCES

Al-Rubaiee, H., Qiu, R., & Li, D. (2016). Analysis of the relationship between Saudi twitter posts and the Saudi stock market. 2015 IEEE 7th International Conference on Intelligent Computing and Information Systems, ICICIS 2015, December, 660–665. https://doi.org/10.1109/IntelCIS.2015.7397193

Alhamad, A., Azis, A. I. S., Santoso, B., & Taliki, S. (2019). Heart Disease Prediction using methods of Machine Learning based on Ensemble – Weighted Vote. 5(3), 352 – 360.

Blatnik, A., Jarm, K., & Meža, M. (2014). Movie sentiment analysis based on public tweets. Elektrotehniski Vestnik/Electrotechnical Review, 81(4), 160–166.

Buntoro, G. A. (2017). Analysis of candidates for governor of DKI Jakarta 2017 on Twitter. Integer Journal March, 1(1),32–41. Https://www.researchgate.net/profile/Ghulam_Buntoro/publication/316617194_Analisis_Sentimen_Calon_Gubernur_DKI_Jakarta_2017_Di_Twitter/links/5907eee44585152d2e9ff992/Analisis-Sentimen-Calon-Gubernur-DKI-Jakarta-2017-Di-Twitter.pdf

Cureg, M. Q., De La Cruz, J. A. D., Solomon, J. C. A., Saharkhiz, A. T., Balan, A. K. D., & Samonte, M. J. C. (2019). Sentiment analysis on tweets with punctuations, emoticons, and negations. ACM International Conference Proceeding Series, Part F1483(1), 266–270. https://doi.org/10.1145/3322645.3322657

Da Silva, N. F. F., Hruschka, E. R., & Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. Decision Support Systems. https://doi.org/10.1016/j.dss.2014.07.003

Flux, A. W., Pareto, V. (1897). Political Economy Course. The Economic Journal. https://doi.org/10.2307/2956966

Gorunescu, F. (2011). Data mining Concepts, Models, and Techniques. Verlag Berlin Heidelberg: Springer

Han, J., & Kamber, M. (2007). Data mining Concepts and Techniques. Morgan Kaufmann publisher.

Jiawei Han, & Kamber, M. (2013). Data Mining: Concepts and Techniques Second Edition. In Morgan Kaufmann. https://doi.org/10.1017/CBO9781107415324.004

Larose, D. T. (2005). Discovering Knowledge in Data. New Jersey: John Willey & Sons, Inc.

Kartiko, M., & Sfenrianto. (2019). Accuracy for Sentiment Analysis of Twitter Students on ELearning in Indonesia using Naive Bayes Algorithm Based on Particle Swarm Optimization. Journal of Physics: Conference Series, 1179(1). https://doi.org/10.1088/1742-6596/1179/1/012027

Mentari, N. D., Fauzi, M. A., & Muflikhah, L. (2018). 2013 curriculum sentiment analysis on Twitter social Media using the K-Nearest Neighbor method and the Feature Selection Query Expansion Ranking. Journal of Information Technology and Computer science development (J-Ptiik) Universitas Brawijaya, 2(8), 2739 – 2743.

Pratama, B., Saputra, D. D., Novianti, D., Purnamasari, E. P., Kuntoro, A. Y., Hermanto, Gata, W., Wardhani, N. K., Sfenrianto, S., & Budamsono, S. (2019). Sentiment Analysis of the Indonesian Police Mobile Brigade Corps Based on Twitter Posts Using the SVM and NB Methods. Journal of Physics: Conference Series, 1201(1). https://doi.org/10.1088/1742-6596/1201/1/012038

Puyalnithi, T., V, M. V., & Singh, A. (2016). Comparison of Performance of Various Data Classification Algorithms with Ensemble Methods Using Rapidminer. 6(5), 1–6.

Rachmat, A., & Lukito, Y. (2016). Implementation of WEB based Crowdsourced Labelling system with Weighted Majority Voting method. ULTIMA Infosys Journal, 6(2), 76 – 82. https://doi.org/10.31937/si.v6i2.223

Ratul, A. R., & Engineering, F. (n.d.). A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining.

Witten, I. H., Frank, E., & Hall, M. a. (2011). Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). In Complementary literature None. http://books.google.com/books?id=bDtLM8CODsQC&pgis=1