SENTIMENT ANALYSIS ON TWITTER USER OF STANDUP COMEDY EVENT BY USING SUPPORT VECTOR MACHINE AND NAÏVE BAYES CLASSIFIER

Rina Priyani

Informatic Engineering – **Indonesian Computer University**
Jl. Dipatiukur 112-114 Bandung
E-mail : rinapriyani27@gmail.com

ABSTRACT

One event is a standup comedy standup comedy KompasTV. The existence of an official account on twitter KompasTV standup comedy, allowing people who become followers of the account do feedback by responding to every posting a tweet from the account. Because of the availability of sentiment for a standup comedy show KompasTV very abundant, so the feedback submitted by followers can be used to measure sentiment analysis followers to standup comedy show in KompasTV. The number of tweets that get into standup comedy KompasTV twitter account can be used to determine the sentiment of followers to show KompasTV standup comedy, but in the delivery of these sentiments are sentiments which is not an opinion, and therefore before sentiment classified into positive and negative categories, a Data twitter should be separated into categories relevant and irrelevant by using Support Vector Machine. Once the data twitter entry into the relevant category, then the data twitter classified into positive and negative categories using the Naive Bayes classifier.

Problems sentiment analysis is how to classify sentiment into classes that have been determined. Furthermore, in order to overcome these problems is done by using a machine learning Support Vector Machine for separation of data relevant and irrelevant, and Naïve Bayes classifier for data classification positive and negative. The use of Support Vector Machine and Naïve Bayes classifier because performasinya better for Indonesian-language text, so the sentiment analysis can be applied to tweet which has a subjective value.

From the test results by the method of 10-fold cross validation showed that both methods used have a satisfactory level of accuracy is 90.74% for the method of Support Vector Machine and 95.91% for the Naïve Bayes classifier method.

Keyword: Analisis sentimen, *Support Vector Machine*, *Naïve Bayes Classifier*, *Cross Validation*, Twitter, *tweet*

1. INTRODUCTION

Nowadays, the development of television Indonesia is very fast, so many emerging new events that varies. Ranging from news shows, talk shows, variety shows up comedy show or a comedy. One of standup comedy is comedy, a genre that is one profession that pelawaknya droll jokes brought on stage by myself, in a way monologue about a topic. People who do this activity called a stand-up comedy or comic [1].

One television that has a standup comedy show that KompasTV. Not just provide entertainment comedy with existing comics, KompasTV had several times held a talent show standup comedy, whose audition spread across Indonesia. Therefore enthusiastic people large enough to standup comedy, then KompasTV also have social media as a promotional tool standup comedy show, one that is twitter. Twitter is used KompasTV for displaying schedules standup comedy show that will be aired so that people can watch the event. The existence of standup comedy KompasTV official account on twitter, allowing people who become followers of the account do feedback by responding to each post a tweet from the account. Due to the availability of sentiment for a standup comedy show KompasTV very abundant, so the feedback given by the followers can be used to measure sentiment analysis followers to standup comedy show in KompasTV.

The problem is how to perform sentiment analysis on the data classification twitter sentiment that the results can be used to evaluate the quality of a television show. The use of Support Vector Machine and Naïve Bayes classifier is due to better results for text classification Indonesian language, so the sentiment analysis can be applied to tweet Indonesian who has a subjective value [2].

2. SENTIMENT ANALYSIS ON TWITTER USER OF STANDUP COMEDY EVENT BY USING SUPPORT VECTOR MACHINE AND NAÏVE BAYES CLASSIFIER

KompasTV a private television station which has a standup comedy programs, and using Twitter as a social media for promotion. With the standup comedy KompasTV account and the number of posts a tweet that many every day there is feedback from the audience allows standup comedy show which is the followers of the twitter account, which can be considered as the availability of sentiment that is used for quality evaluation program KompasTV standup comedy.

But the problems in the sentiment analysis is how to classify sentiment on twitter of data that results can be used to evaluate programs in KompasTV standup comedy.

2.1 Analysis System

In implementing the method of Support Vector Machine and Naïve Bayes classifier will be built a prototype with a description of the system which will be shown in Figure 1.

The system will be built has four stages: preprocessing, the data training, testing, and cross validation. For an explanation of each step can be seen as follows:

1. The first stage is the stage of retrieving data using twitter by crawling stream API, then the data is done preprocessing process which is the stage where the data will be used as a data testing cleared of noise or from something that has no influence on sentiment. Preprocessing is done in this phase include: case folding, cleansing, stopword removal, convert emoticons, convert negation, tokenizer.
2. The second stage is the stage of training the data, at this stage uses two training data, the training data support vector machine and the training data naïve Bayes, then the training data is done preprocessing stage and classified manually into four classes, namely classes relevant and irrelevant to training data support vector machine, and positive and negative grade for the training data naïve Bayes. Then all four classes are stored into a database of training data. After the tweets that have been classified manually created models sentiment analysis that will be saved back into the database detail the training data that will be used in the testing stage as sentiment analysis model.
3. The third stage is the stage of testing, at this stage using support vector machine testing of data and data testing naïve Bayes as a data input which of course is through the preprocessing stage. Both testing the data is processed so that it can be seen class sentiments.
4. The fourth stage is the stage of testing using the method of k-fold cross validation in which the data is divided number of n-fold is desired, then the process of testing and training conducted n times until the results come out accuracy. The accuracy of the results obtained from the average accuracy at each iteration.

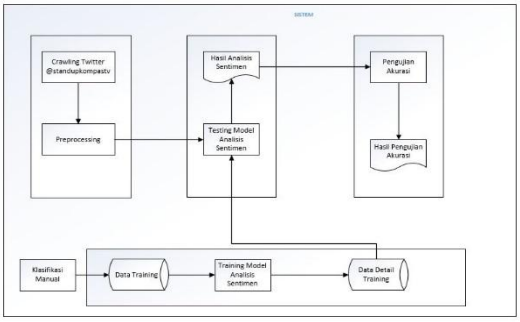


Figure 1. System overview

2.2 Analysis of Data Input

The input data used is the tweet from the official Twitter account standup KompasTV. Tweet data is obtained by utilizing the features of API (Application Interface) that has been provided by twitter for mendapakan API key, and access token that will be used for data retrieval tweet. The data is data taken tweets that contain the word "standupkompastv" tweet or data contained in the official account "StandUpKompasTV", the data can be considered to represent the sentiment of the user or standup KompasTV followers. Data in the form of a sentence with a maximum length of 140 characters. Examples of every tweet containing sentiment on account "StandUpKompasTV" contained in Figure 2 and Figure 3.



Figure 2. Tweets with positive sentiment



Figure 3. Tweets with negative sentiment

Tweet Data taken from the official account KompasTV Standup comedy is still a raw data and there is noise or traits that have no effect on sentiment classification as a link, "@", stopword, which is marked by the hashtag "#" character. Examples of such raw tweet exist in Figure 4.



Figure 4. Example of data tweet

In addition, the tweet perceived characteristics have an influence on the determination of the sentiment is there an emoticon that is one way to express approval or disapproval phrase in a sentence or tweets. Examples tweet existing of crude emoticons contained in Figure 5.



Figure 5. Example of data tweet with emoticons

Characteristics next tweet is no negation there is a word that needs to be considered in the analysis because the sentiment can change the sentiment of a tweet. The word is the negation as "tidak", "bukan", "ga", "jangan", "nggak", "tak", "tdk", dan "gak". For example the word "lelet" which is a word with negative sentiment, because in front of the word "lelet" there is a negation word "tidak" then the sentiment of the combined word is positive. Examples tweet contained a negation word is in Figure 6.



Figure 6. Example of data tweet with negation

Tweets so that the data can be put to good use to classify sentiment will require preprocessing process. In this process the data tweet that used for testing the data the data training data dibersikan of noise or traits that have no effect on sentiment classification as a link, "@", "RT", stopword, and so forth. Preprocessing process description can be seen in Figure 7.

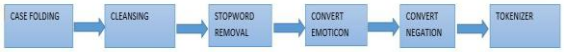


Figure 7. Stages Preprocessing

1. *Case Folding*

In the case of folding process large or uppercase letters contained in a tweet converted to lowercase.

2. *Cleansing*

Tweet contained in the official account standup KompasTV has various components or characteristics tweet characteristic such as "@" are identified as components of usernames, URLs, known through the operation of regular, hashtag which marks the word as a topic that is being discussed, and "RT" are identified as repeat the tweet was posted. These components do not have any influence on the sentiment, it will be discarded.

3. *Stopword Removal*

Data tweets that have been through the previous process still contains words deemed unable to give effect to determine a category sentiment. The words are inserted into the stopword list which is usually in the form of personal pronouns, relative pronoun, pronomial pointer, and others.

4. *Convert Emoticon*

Convert emoticon is the process of converting emoticon into a string corresponding to the emoticon expression itself. Convert emoticons made because the tweet of data taken from standup comedy emoticon KompasTV there is one way of expressing approval or disapproval in a tweet. It is considered to have an influence on sentiment classification, therefore convert emoticons used.

Table 1. Conversion Emoticon[3]

Emoticon	**Conversion**
>:] :-) :) :o) :] :3 :c) :> =] 8) =) :} :^)	ehappy
>:D :-D :D &-D 8D x-D xD X-D XD =-D =D =-3 =3	elaugh
>:] :-(:(:-c :c :-< :< :-[:[:{ <><<><	esad
D:< D: D8 D; D= DX v.v D-‘:	ehoror
>:P :-P :P X-P x-p xp XP :-p :p =p :-b :b	etongue
>:o >:O :-O :O o_O o.O 8-0	eshock
>:\ >:/ :-/ :-. :/ :\ =/ =\ :S	eangry
:| :-|	eflat

5. *Convert Negation*

Convert negation is the process of converting the words of negation contained in a tweet, because the word negation have an influence in changing the value of sentiment in a tweet. Said negation contained in a tweet would be eliminated, and given marker. If there is a negation word will eat united with said afterwards. The words of negation include words "bkn", "bukan", "tida", "tak", "ga", "enggak", "g", "jangan", dan "tidak".

6. *Tokenizer*

In the process tokenizer every word on the tweet is separated, in the process steps done is to separate each word separated by a space, the next part of tweets that have a character other than the alphabet, numbers, and the bottom line will be broken down according to the position the character and the part that has only one non-alphabetic and numeric characters will be discarded.

2.3 Term Weighting

Term weighting is the process giving weight to every word in the document, the term weighting method used in this study is the TF IDF. Term frequency is one of the simplest methods of weighting. In this method, each term is assumed to have the proportion of interest in accordance with the number of occurrence (emergence) of the terms in the document. TF equation is as follows:

$$W\ d,t\ = TF(d,t)$$

Where TF (d, t) is the frequency of occurrence of term t in document d.

Inverse Document Frequency attention to the appearance of the term in the document. In this method, a term that is considered valuable is the term that rarely appear in the document. IDF equation is as follows:

$$IDF\ t\ = log\frac{N}{df(t)}$$

Where df (t) is many documents that contain the term t. TF * IDF is a combination of TF with the IDF. So the equation TF * IDF is as follows:

$$TF * IDF\ d,t\ = TF\ d,t\ * IDF(t)$$

2.4 Support Vector Machine

Support Vector Machine (SVM) is a machine learning method that converts text into vector data. Vector in this study has two components, namely the dimension (word id) and weights. SVM concepts can be explained simply as an attempt to find the best hyperplane which serves as a separator are two classes in the input space[4]. Attempt to find the location of this hyperplane is the core of the learning process on SVM. SVM classification is divided into two, namely linear and nonlinear.

Starting with the linear classification of cases, this function can be defined as.

$$g\ x\ := sgn(f\ x\ )$$

with $f\ x\ = wx + b$

$$\text{Or } g\ x\ = \begin{matrix} +1, jika\ \ wx + b\ \geq +1 \\ -1, jika\ \ wx + b\ \leq -1 \end{matrix}$$

Input to the SVM research consists of the points which is a vector of real numbers available data is denoted as xi ∈R ^ d while each label is denoted as yi∈ {-1, + 1} for i = 1, 2, ..., 1 where 1 is the number of data. Both classes are assumed to -1 (negative) and +1 (positive) can be separated perfectly by hyperplane dimension d, which is defined,

$$\text{with } wx + b = 0$$

A pattern xi which belongs to the class -1 (negative samples) can be formulated as a pattern that satisfies the inequality:

$$\text{With } wx + b \leq -1$$

While the pattern that belongs to the class xi +1 (positive samples):

$$\text{With } wx + b \geq +1$$

The margin can be found by maximizing the value of the distance between the hyperplane and the closest point, which is 1 / (‖w‖).

2.5 Classification with Naive Bayes

In using the naïve Bayes classification is divided into two processes, namely the process of training and testing. The training process is used to produce a sentiment analysis model that will be used as a reference to classify sentiment data testing of new or raw data. Here is sentiment classification algorithms using Naïve Bayes Classifier:

1. Training
   a. Calculate p (c_i).
   b. Calculate p $w_k\ c_i$)for each word $w_k$on the model.
2. Testing
   a. Calculate $_k$p w_k c) × p (c) for each category
   b. Determine the value category $_k$p w_k c) × p c maximum.

Determination of the category of a document is done by choosing the value of *c* which has a value of $p(C = c_i | D = d_j)$ maximum, and is expressed by:

$$c^* = arg\ max_{c\ \in C}\ \ _k p\ \ w_k\ c\ ) \times p\ (c) \quad (1)$$

Where :

$_k p\ \ w_k\ \ C = c_i)$ is the product of the probability of occurrence of all words in the document*dj*.

$p\ \ c$ is the result of the probability of all categories in the training data.

Category*c** is the category that has the value of$p(C = c_i \mid D = d_j)$ maximum. Value of$p(D = d_j)$ does not affect the comparison because for each k value will be the same category.

3. IMPLEMENTATION AND TESTING

3.1 Environment Implementation

The hardware used in the design and implementation of the system has some specifications as follows:

1. Processor 2.13GHz
2. RAM 1024MB
3. Monitor 1366x768 pixels
4. Keyboard
5. Mouse

The software used in the implementation process of this system are as follows:

1. Windows Operating System 8 Starter 32 bit

2. PHP programming language.
3. Web server XAMPP.
4. Code editor berupa SublimeText3Portable.
5. DBMS : MySQL 5.5.27
6. Web Browser: Google Chrome 34.0.1847.137

3.2 Implementation of Interfaces

In the sub-chapter will explain the implementation of the interface has been designed previously. Results of interface design that has been implemented is as follows:

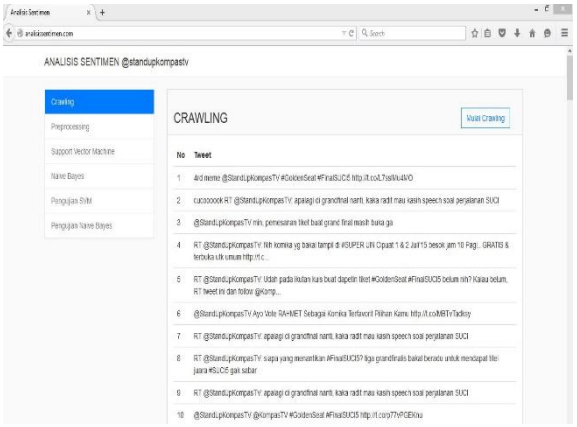


Figure 8. Interface of Crawling Tweet

Figure 8. Tweet Crawling is an interface that is implemented in the system. This tab serves for crawling the data tweet from the official account KompasTV standup comedy, which will be used as a data testing to be classified into several categories sentiment.

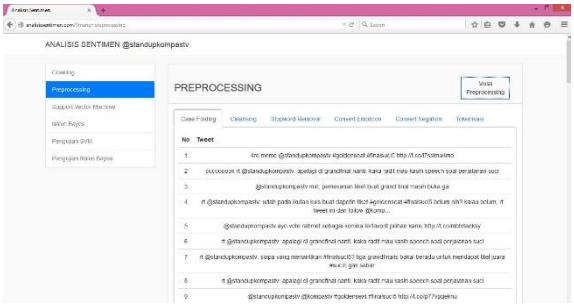


Figure 9. Interface of *Preprocessing*

Figure 9. a interface of preprocessing that serves that function for displaying the results of any testing preprocessing of data before the data is in klasfikasi, such as case folding, cleansing, stopword removal, convert emoticons, convert negation, tokenizer.

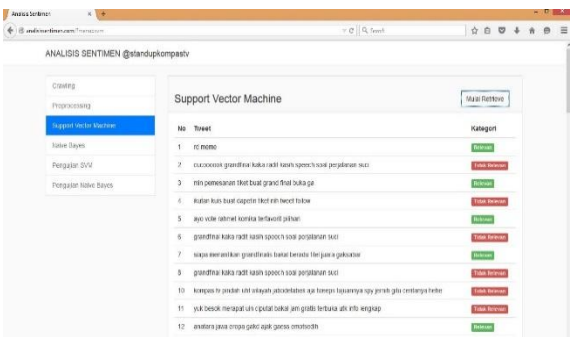


Figure 10. Interface of Support vector machine

Gambar 10. a interface of support vector machine that serves to display the data tweet that has been grouped into categories relevant and irrelevant.

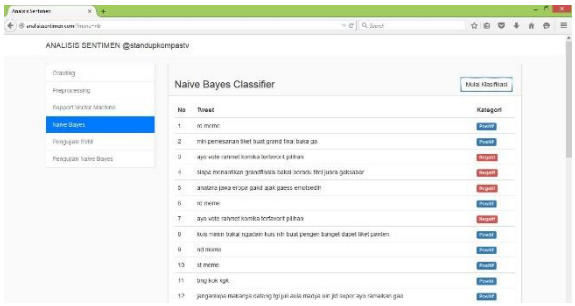


Figure 11. Interface of Naïve Bayes

Figure 11. a interface of naïve Bayes that serves to display the data tweet that has been classified into positive and negative categories. the results of testing or the results of testing data classification, sentiment classification of each tweet on the data obtained from the testing naïve Bayes calculations that had previously been through the preprocessing stage.

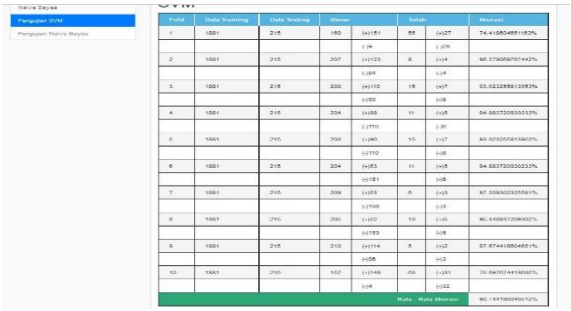


Figure 12. Interface of Testing Support Vector Machine

Figure 12. a interface of testing support vector using cross validation that serves to test the performance of the classification system in the wake.

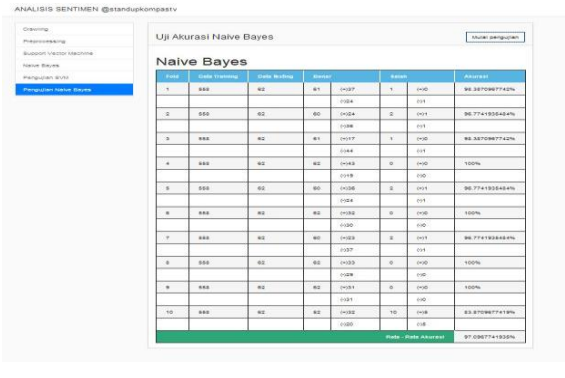


Figure 13. Interface of Testing Naïve Bayes Classifier

Figure 13. a interface of testing naïve using cross validation that serves to test the performance of the classification system in the wake. Performance classification can be seen from the calculation accuracy by cross validation method.

3.3 Accuracy Testing Method System with 10-fold cross validation

The amount of data plays an important role in machine learning algorithms. The amount of data bit (<100 instance) may make a machine learning algorithm is not accurate. Machine learning algorithms recommend the number of instances that many (> 1000 instance) but the data itself is not easy to obtain.

K-fold cross validation is a technique that can be used if it has a limited amount of data (number of instances not many) [5]. K-fold cross validation is one method used to determine the average success of a system by means of looping to randomize the input attributes that the system is tested for several attributes of random input. K-foldcross validation begins by dividing the data number of the desired n-fold. In the process of cross validation data will be divided into n partitions of the same size D1, D2, D3..Dn further testing and training processes carried out n times. In the i-th iteration will be a data partition in testing and the remainder will be the training data. To use the best fold number for validity, it is recommended to use 10-fold cross validation in the model.

Scenario testing is determining the stage of testing. Tests conducted for the method of support vector machine and naïve Bayes. Testing is done using k-cross validation with k values as much as 10 fold, this test aims to determine the accuracy of the method of support vector machine and naïve Bayes classifier that is applied to the analysis of sentiment when tested with the training data and testing the data differently. The testing phase using 10-fold cross validation divides the dataset which initially amounted to 1881 data will be divided into 10 subsets of data for testing support vector machine. As for the naïve Bayes classifier testing phase using 10-fold cross validation divides the dataset that originally numbered 549 data will be divided into 10 subsets of data. At first fold there is a combination of 9 different subsets combined and used as training data, while the first subset (remainder) is used as a data testing, further training and testing process is done until the tenth fold. Scenario test the accuracy of the method 10-fold cross validation can be seen in Table 2.

Table 2. Stability Test Scenario with 10-fold cross validation

Fold	**Data**	***Subset***
Fold 1	Training Testing	$S_2, S_3, S_4, S_5, S_6, S_7 S_8, S_9, S_{10}$ S_1
Fold 2	Training Testing	$S_1, S_3, S_4, S_5, S_6, S_7 S_8, S_9, S_{10}$ S_2
Fold 3	Training Testing	$S_1,S_2, S_4, S_5, S_6, S_7 S_8, S_9, S_{10}$ S_3
Fold 4	Training Testing	$S_1,S_2, S_3, S_5, S_6, S_7 S_8, S_9, S_{10}$ S_4
Fold 5	Training Testing	$S_1,S_2, S_3, S_4, S_6, S_7 S_8, S_9, S_{10}$ S_5
Fold 6	Training Testing	$S_1,S_2, S_3, S_4, S_5, S_7, S_8, S_9, S_{10}$ S_6
Fold 7	Training Testing	$S_1,S_2, S_3, S_4, S_5, S_6, S_8, S_9, S_{10}$ S_7
Fold 8	Training Testing	$S_1,S_2, S_3, S_4, S_5, S_6, S_7,S_9, S_{10}$ S_8
Fold 9	Training Testing	$S_1,S_2, S_3, S_4, S_5, S_6, S_7$, S_{10} S_9
Fold 10	Training Testing	$S_1,S_2, S_3, S_4, S_5, S_6, S_7 S_8, S_9$ S_{10}

3.4 Testing Results and Analysis

In the test scenario with the number of 10-fold cross validation and iterations were performed 10 times iterations to support vector machine generated average is 90.74%. The first iteration of the first fold using 1881 data, the training process and the process of testing carried out until the 10th iteration. Details of the results of testing the accuracy of scenarios for support vector machine on each foldnya can be seen in Table 3.

Table 3. Results of Stability Test with 10-fold cross validation to support vector machine

Fold	Data Training	Data Testing	Accuracy
1	1881	215	74.42%
2	1881	215	96.28%
3	1881	215	93.02%
4	1881	215	94.89%
5	1881	215	93.02%
6	1881	215	94.89%
7	1881	215	97.21%
8	1881	215	95.35%
9	1881	215	97.67%
10	1881	215	70.70%

Rata-rata akurasi	90.74%

In the test scenario with the number of 10-fold cross validation and iterations were performed 10 times iterations for naïve Bayes resulting average is 95.91%. The first iteration of the first fold using the data 549, the training process and the process of testing carried out until the 10th iteration. Details of the accuracy of the test results for naïve Bayes scenario on each foldnya can be seen in Table 4.

Table 4. Results of Stability Test with 10-fold cross validation to support vector machine

Fold	Data Training	Data Testing	Accuracy
1	549	66	96.97%
2	549	66	95.45%
3	549	66	96.97%
4	549	66	98.49%
5	549	66	95.45%
6	549	66	98.49%
7	549	66	95.45%
8	549	66	98.49%
9	549	66	98.49%
10	549	66	84.85%
Rata-rata akurasi			95.91%

Table 3 indicates that the accuracy of each fold has been very stable. Experiments on fold to-1 shows the results of the lowest among 9 other fold is ≈ 74.42%, while the remaining fold yield greater accuracy with aurasi highest in the fold 1, 2, 3, 5, 7, and 10. For table 4 also indicate that the accuracy of each fold has been very stable. Experiments on fold-10 show lower results among 9 other fold is ≈ 84.85%. When analyzed the performance of the classification with Support Vector Machine and Naïve Bayes produced can be said to be accurate, as well as the performance of the classification model can be said to be good to see an average accuracy of each fold are large enough that ≈ 90.74% for the support vector machine and ≈ 95.91% naïve Bayes for the 10 fold cross validation with training data is used as data for the 1881 and 549 support vector machine data for naïve Bayes, because in general the larger the training data, the better the performance of the classification model of a system.

4 CLOSING

4.1 Conclusion

Of the process of testing and analysis has been done, conclusions can be drawn, among others:

1. Method of Support Vector Machine can be applied to the analysis of sentiment in the retrieval of data to classify the data relevant and irrelevant.
2. The method of Naïve Bayes classifier can be applied to the analysis of sentiment in classifying positive and negative sentiment.
3. The results of stability testing classification performance by using 10-fold cross validation generate sufficient accuracy as high as 90.74% for Support Vector Machine and 95.91% for the Naïve Bayes classifier. This shows that the classification model performance has been quite good and stable.

4.2 Suggestions

From the test results, analysis and conclusions have been formulated, there are things you suggested for further research, the sentiment analysis in this study has not noticed that the semantic meaning of words and sentences. Future studies are expected to use the meaning of words and sentences to determine the category of sentiment on a document in order to obtain better results.

BIBLIOGRAPHY

[1] [Online]. Available : http://suc.metrotvnews.com/article/ensiklopedia/13. [Accessed 12 Februari 2014].

[2] Ni Wayan Sumartini Saraswati, "pps.unud.ac.id," [Online]. Available: www.pps.unud.ac.id/thesis/pdf_thesis/unud-209-236721286-tesis.pdf. [Accessed 10 Februari 2014]

[3] I. Sunni and D. H. Widyantoro, "Analisis Sentimen dan Ekstraksi Topik Penentu Sentimen pada Opini Terhdap Tokoh Publik," *Jurnal Sarjana Institut Teknologi Bandung Bidang Elektro dan Informatika*, vol. 1, pp. 200 - 206, 2012.

[4] Nugroho A.S., Arief Budi, and Dwi Handoko. 2003. Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika, kuliah umum IlmuKomputer.Com.

[5] R. Kohavi, "A study of Cross-Validation and Bootstrap for Accuracy Estimation andModelSelection,"1995. [Online].Available:http://ijcai.org/Past%20Proceedings/IJCAI-95-VOL2/PDF/016.pdf. [Accessed 18 Desember 2014].