Comparison Of Feature Selection For SVM Sentiment Classification On Twitter

Dede Juniawan Suri[1], Ken Kinanti Purnamasari[2]

[1] Informatics engineering – Indonesian Computer University
Jl. Dipatiukur 112-114 Bandung
[2] Informatics engineering – Indonesian Computer University
Jl. Dipatiukur 112-114 Bandung
E-mail : dede.juniawan.suri@gmail.com[1], guinev_gs@yahoo.co.nz[2]

ABSTRACT

Sentiment analysis is the field of study to analyze the opinions, sentiments, evaluation, assessment and emotional attitude towards entities such as products, services, organizations, individuals, events, and other attributes. Feature selection has an important role in the classification of text including sentiment analysis, the use of feature selection methods to help understand the attributes that are relevant for a particular class and to improve classification accuracy. Feature selection methods TF-IDF, Ratio, and N-Gram has been applied in the classification of short message twitter, these methods produce good accuracy, but not known better method because it is done in different research. Based on the results of the study with 1100 tweets dataset is divided into 500 for training tweet filter, 500 for sentiment classification training, 50 for testing filter tweets, and 50 for testing sentiment classification. Methods TF-IDF produce the highest accuracy, with 74% on the classification filter tweets and 80% on the sentiment classification sentiment, Ratio produces accuracy 72% on the classification filter tweets and 74% on the sentiment classification, and N-Gram produce accuracy 64% on the classification filter tweets and 78% on sentiment classification.

Keyword: Sentiment Analysis, Feature Selection, TF-IDF, Ratio, N-Gram, Support Vector Machine, Twitter

1. INTRODUCTION

Sentiment analysis is the field of study to analyze the opinions, sentiments, evaluation, assessment and emotional attitude towards entities such as products, services, organizations, individuals, events and other attributes. Sentiment analysis focuses on the opinions expressed positive or negative sentiments. Selection feature has an important role in the classification of text including sentiment analysis, the use of feature selection methods to help understand the attributes that are relevant for a particular class and to improve classification accuracy.
Feature selection methods TF-IDF, Ratio, and N-Gram has been applied in the classification of short message twitter and produce good accuracy. However, there are problems to determine the method of selection of the better features because it is done in different research with the use of the number and different data sources, both in the process of training and testing, and therefore to determine the method of selection of the better features for sentiment analysis should be conducted a comparison for each feature selection method ever applied to the source or the same amount of data.

1.1 Sentiment Analysis

Sentiment analysis or opinion mining is a field of study to analyze the opinions, sentiments, evaluation, assessment and emotional attitude towards entities such as products, services, organizations, individuals, events and other attributes. Sentiment analysis focuses on the opinions expressed positive or negative sentiments.

There are three general stages in the process of sentiment analysis, which is taking the data from the data source or so-called crawling, the next phase of training and testing using a specific classification algorithm. One algorithm to perform the classification is Support vector machine. The source of data that can be used to process one is sentiment analysis of social media Twitter. Twitter as one of the interactive social network allows users to give an opinion on a matter in real time. The data can be analyzed on Twitter that tweet created by users, where these tweets will be classified into two classes, namely sentiment sentiment is positive or negative sentiment.

1.2 Seleksi Fitur

Feature selection as a stage in the classification is divided into two approaches, namely wrapper and filter approach. Feature selection filter approach is carried out separately by classification engine, or in other words be a feature selection stage of the pre-process before the data is loaded into the machine classification. A data feature selection has been done with this approach, it can be used or combined with some classification engine. In contrast to the filter

approach, the approach wrapper to make classification engine as part of an important process in determining whether or not a feature. Machine classification will be trained and tested with a combination of features that change, if the addition of a feature lowers machining accuracy classification feature, it is not used. The machine will process a combination of other features and do not use features that will weaken the system's accuracy.
This research will apply the filter feature selection approach because it is simpler than the wrapper approach. In addition, the time required to filter approach is faster than the wrapper approach. As for feature selection method used, namely TF-IDF, Ratio, and N-Gram.

1.1.1 TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is method of term weighting. Term weighting is the process giving weight to all words in the document.
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 term in the document. TF equation is as follows.

$$W(d,t) = TF(d,t) \quad (1)$$

Where TF (d, t) is the frequency of occurrence of the term t in the document d.
Inverse Document Frequency attention to the appearance of the term in the document. In this method, a term that is considered valuable / precious are terms which rarely appear on the collection / document. IDF equation is as follows.

$$idf(t) = log\frac{N}{df(t)} \quad (2)$$

Where df (t) is a lot of documents that contain the term t.
TF-IDF is a combination of TF with IDF method. TF-IDF so the equation is as follows.

$$TF - IDF(d,t) = TF(d,t).IDF(t) \quad (3)$$

1.1.2 *Ratio*

Ratio method is based on Information Theory, in this method, an increase in the value or weight of the features when word often occurs in a classroom and will decrease when the word occurs in all classes. The steps in this method are as follows.

1. Calculate value tf_i

$$tf_i = \sum_{j=1}^{n} F_{(i,j)} \quad (4)$$

2. Calculate the value FR (Frequency Ratio), this will give you an idea how important the word i to j group elected

$$FR_{(i,j)} = \frac{F_{(i,j)}}{tf_i} \quad (5)$$

3. Calculate MRF (Maximum Frequency Ratio)

$$MRF_i = max\{FR_{(i,j)}\} \quad (6)$$

1.1.3 N-Gram

One of method that can be used to take on the features that the N-Gram tweet. N-gram pieces of n-characters taken from a string. N-Gram method used to take pieces of code letters in a word number n is continuously read from said source to the end of the document. If n = 1 then called Bigram, called Bigram n = 2 and n = 3 is called trigrams.
Advantage N-Gram is based on the characteristics of the N-Gram as part of a string, so that errors in some strings will only result in a difference in the partial N-Gram. For example, if the N-Gram of the two strings are compared, then calculated the same N-Gram of two strings that it will get the value of similarity or similarity of the two strings that are resistant textual errors.
For example, the similarity between the raw word "cool" and said no default "crane" can be measured by calculating the similarity to how many pieces of n-grams derived from the two words have in common, namely:

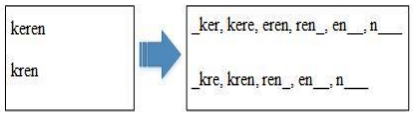


Figure 1 Examples of N-Gram

N-Grams are raised on those two words, there are 3 n-gram similarity of the two words. Visible there are only two parts of n-gram in the word "kren" are not contained in the word "keren", so the word "kren" can be assumed to be "keren".
In Johannes study [4] use of the feature 2-gram and 3-gram helpful for increased performance categorization. On the other hand, the use of n-grams degan n> 3 instead reduce the performance of categorization, so in this study will apply the N-Gram with n = 3 or trigram.

1.3 *Support vector machine*

Support vector machine (SVM) is a machine learning method that converts text into vector data. Vector in this study had two components, 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. [5]

In a linear support vector machine separator is a linear function. Training data is expressed by (x_i, y_i) and x_i = { x_1, x_2, …,x_{iq}} are attributes (features) for the training data set to the class i. For y_i € {-1,1} stated label. defining equation of a hyperplane dividing written by:

$$wx_i + \mathrm{b} = 0 \tag{7}$$

Data x_i are divided into two classes, which belongs to the class -1 (negative sample) is defined as a vector that satisfies the following inequality 8.

$$wx_i + \mathrm{b} < 0 \; untuk \; y_i = 1 \tag{8}$$

While belonging to a class +1 (positive samples) satisfy the following inequalities 9.

$$wx_i + \mathrm{b} > 0 \; untuk \; y_i = +1 \tag{9}$$

Where:

x_i = data input

y_i = label given

w = value of the normal field

b = position of the field relative to the origin

The parameters w and b are parameters to be sought value. When the data labe y_i = -1, then the barrier into the equation 10 below.

$$wx_i + \mathrm{b} \leq -1 \tag{10}$$

When the data label y_i = +1, then the barrier into the equation 11 below.

$$wx_i + \mathrm{b} \geq +1 \tag{11}$$

The margin can be searched by maximizing the distance between the two classes and field delimiter closest point, namely $\frac{2}{|w|}$. It is formulated as a problem of quadratic programming (QP) problem is to find the point of at least equation (12) by taking into account equation (13) below.

$$\min \tau(w) = \frac{1}{2}||w||^2 \tag{12}$$

$$y_i(wx_i + b) - 1 \geq 0, (i = 1, \dots, n) \tag{13}$$

This problem can be solved by sharing computational techniques. More easily solved by changing the equation 12 into the equation 14 Lagrangian function, and reduce it to 15 the following equation.

$$L(w, a, b) = \frac{1}{2}\left|\left|w\right|\right|^2 - \sum_{i=1}^{n} a_i(y_i((w^T x_i + b) - 1)) \tag{14}$$

$$L(w, a, b) = \frac{1}{2}\left|\left|w\right|\right|^2 - \sum_{i=1}^{n} a_i \, y_i(w^T x_i + b) + \sum_{i=1}^{n} a_i \tag{15}$$

Where a_i lagrange multiplier is zero or positive ($a_i \geq 0$). The optimal value of the equation 15 may be calculated by minimizing L to w, b, and a. Can be seen in equations 16 through 18 below.

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} a_i \, y_i x_i = 0 \tag{16}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} a_i \, y_i = 0 \tag{17}$$

$$\frac{\partial L}{\partial a} = w - \sum_{i=1}^{n} a_i \, y_i(w^T x_i + b) - \sum_{i=1}^{n} a_i = 0 \tag{18}$$

Then the Lagrange problem for classification can be expressed in the following equation 19.

$$Min \; L(w, a, b) = \frac{1}{2}\left|\left|w\right|\right|^2 - \sum_{i=1}^{n} a_i \, y_i(w^T x_i + b) + \sum_{i=1}^{n} a_i \tag{19}$$

Having regard to the equations 20 and 21 below.

$$w - \sum_{i=1}^{n} a_i \, y_i x_i = 0 \tag{20}$$

$$\sum_{i=1}^{n} a_i \, y_i = 0 \tag{21}$$

Equation 19 is a model of primal Lagrange. While maximizing L against a_i, it is become equation 22 below.

$$Max \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1,j=1}^{n} a_i y_i {y_j}^T \, x_i {x_j}^T \tag{22}$$

Having regard to the following equation 23.

$$\sum_{i=1}^{n} a_i y_i = 0, a_i > 0 \; (i, j = 1, \dots, n) \tag{23}$$

To find the value of x_i and y_i to do when it obtained the value of each word (term) of the weighting of each

feature selection method and class initialization. Results of weighting converted into SVM's format (x), while the class data into a data label SVM (y).
To get the value a_i, the first step is to turn every tweet into a vector value (support vector) = $\binom{x}{y}$. Then the value of each tweet vector inserted into equation 24 kernel trick phi (φ) below.

$$\varphi\begin{bmatrix}x\\y\end{bmatrix} = \begin{cases}\sqrt{x_n^2+y_n^2} > 2, maka \begin{bmatrix}\sqrt{x_n^2+y_n^2}-x+|x-y|\\ \sqrt{x_n^2+y_n^2}-x+|x-y|\end{bmatrix} \\ \sqrt{x_n^2+y_n^2} < 2, maka \begin{bmatrix}x\\y\end{bmatrix}\end{cases} \quad (24)$$

Values of x obtained from the equation 25 linear kernel for x follows.

$$\sum_{i=1,j=1}^{n} x_i^T x_j \, (i,j=1,\dots,n) \quad (25)$$

Values of y obtained from the equation 26 to y the following linear kernel.

$$\sum_{i=1,j=1}^{n} y_i^T y_j \, (i,j=1,\dots,n) \quad (26)$$

To get optimum distance perpendicular to consider the vector is positive, then the calculation of the substitution values of x and y values into the equation 24 rated bias 1. Then search parameters a_i, by first searching for the value of the function of each tweet using the equation 26, and then look for value a_i on linear equations using the equation 27 with regard $i,j=1,\dots,n$ the following.

$$\sum_{i=1,j=1}^{n} a_i {T_i}^T T_j \quad (27)$$

$$\sum_{i=1,j=1}^{n} a_i {T_i}^T T_j = y_i \quad (28)$$

After a_i parameters obtained, then enter into the equation 29 below.

$$W = \sum_{i=1}^{n} a_i T_i \quad (29)$$

Results obtained using the equation 29, then used the equation 30 to obtain the value of w and b is as follows.

$$y = wx + b \quad (30)$$

Such that the obtained value w and b value or value hyperplane to classify both classes.

2. CONTENTS OF RESEARCH

2.1 System Architecture

The system built has several stages. These stages are tweet-making process, the training process, the process of testing and the accuracy of the counting process. A general description of the system built is shown in the following figure.

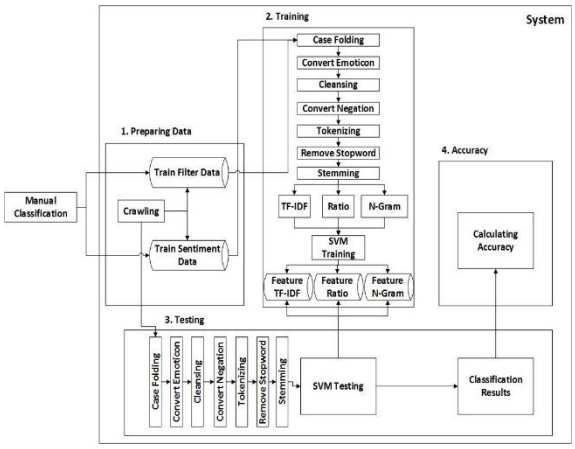


Figure 2 System Overview

The following description of the figure 2.

1. First part is the process of decision-tweet (crawling) is then stored in the data table filter training and the training data and the sentiment classification process is done manually.
2. Second part is the SVM training process, in this study the training is divided into two parts, namely training tweet containing sentiment classification and contain no sentiment (sentiment/not sentiment) then training sentiment classification. But before the training data through the first preprocessing stage, the stage includes a folding case, convert emoticons, cleansing, convert negation, tokenizing, remove stopword, stemming, and then feature selection with TF-IDF method, Ratio, and N-Gram.
3. Third part is the process of testing the SVM. Tweet or testing the data must first go through the preprocessing stage (such as preprocessing stage in the second part), then do the matching process token preprocessing stage results with the features available in the database of the results of training (part 2), the result is a sentiment tweet or not sentiment. Tweet classified as sentiment will be input testing sentiment classification.
4. Fourth section is the accuracy of the calculation process, this step is done to see the effect of feature selection method with the accuracy of the classification.

2.2 Data Input

The input data used is the tweet from twitter account Net TV. The tweet Data obtained by utilizing the features of the API (Application Programming Interface) provided by Twitter. Data taken tweet is data available on Net TV and SCTV account or that contains keywords @netmediatama, #nettv,

#netmediatama, and @SCTV_. The data obtained in the form of text with a maximum length of 140 characters and still be raw data and there is a noise that does not have an influence on sentiment classification as a URL, hashtag (#), username (@), and affix necessitating preprocessing stage.

2.3 Analysis Preprocessing Stage

Based on the analysis of input data that has been described, it is necessary to do the preprocessing stage. Preprocessing stage that will be done is case folding, convert emoticons, cleansing, convert negation, tokenizing, remove stopword, stemming, and feature selection. Overview of the preprocessing stage can be seen in the following figure.

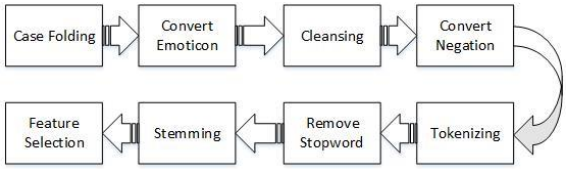


Figure 3 Stage of Preprocessing

1. Case Folding

Case Folding is the uniform form of words on a tweet to lowercase (lowercase) or capital letters (uppercase). In this resaerch, homogenized into lowercase.

2. Convert Emoticon

Emoticons is one of the ways of expressing feelings textual expression. This will assist in determining sentiment in a sentence or tweet. [9] Each emoticon will be converted into the corresponding string, this process is called convert emoticon.

3. Cleansing

Cleansing is the process of cleaning the words that are not required to reduce noise. The words were omitted that URL, hashtag (#) and username (@). It also punctuation such as periods (.), Comma (,) and other punctuation will be omitted because it does not have any effect on the calculation of the value of sentiment. [9] If there is a number, it will be eliminated. In addition, if there is a space of more than one, it will be changed into one space.

4. Convert Negation

A tweet created by users, allowing there is negation to reverse the meaning of a word. It is necessary to convert the stage of negation that by combining the negation word by word thereafter.[10]

5. Tokenizing

Tokenizing is to separate words in tweets into single words. The separation marked with a separator is a space.[9]

6. Remove Stopword

Remove stopword is the process of removing the words contained in stopword list. The words that are in the list stopword form of the personal pronoun, pronoun and bookmark.[9]

7. Stemming

Stemming is a process that transforms words into words berimbuhan essentially (root word). [11] For example, the word fun, please, called the same can be said or one group and can be represented by a common word happy. The process of stemming the Indonesian-language text differs from stemming the English text. In the English text, a process that takes only a process of removing the suffix. While the Indonesian-language text, besides suffixes, prefixes, and konfiks also eliminated. In this study, the algorithm used in the process of stemming namely Nazief and andriani

8. Feature Selection

This resaerch using three methods of feature selection that to be compared to determine the method of selection of the best features for sentiment analysis on Twitter, third adapan the feature selection methods that TF-IDF, Ratio, and N-Gram with the number n=3.

2.4 Analysis of Phase Classification

Tweet classification stage is divided into two stages, namely filter tweets (sentiment and not the sentiment) and the classification of sentiment (positive and negative). Tweets that do not contain the elements necessary to filter these sentiments in advance in order to produce good accuracy on sentiment classification stage. The classification method used in this study is the support vector machine.

2.5 Testing

Testing accuracy of this research is to measure the level of accuracy of each method of selection to the classification by using Support Vector Machine.
The amount of data that is used as much as in 1100 tweet twitter account taken of Net TV and SCTV. The data is divided into 500 data for training the filter tweets, 500 training data for sentiment classification, 50 data for testing filter tweets, and 50 data for testing sentiment classification with the spread of the data as follows.

Tabel 1 Data Of Filter Tweet Classification

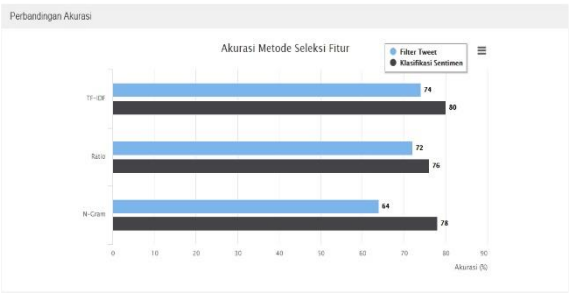
	Not Sentiment	Sentiment	Total
Training Data	250 *tweet*	250 *tweet*	500 *tweet*
Testing Data	22 *tweet*	28 *tweet*	50 *tweet*

Tabel 2 Data of Sentimen Classification

	Negative	Positive	Total
Training Data	232 *tweet*	268 *tweet*	500 *tweet*
Testing Data	20 *tweet*	30 *tweet*	50 *tweet*

Tabel 3 Test Result

	Klasifikasi *Filter Tweet*	Klasifikasi Sentimen
TF-IDF	74%	80%
Ratio	72%	76%
N-Gram	64%	78%



Gambar 4 Graph of Test Result

3. CLOSING

3.1 Conclusion

Based on the results of the research with 1100 tweets dataset is divided into 500 for training tweet filter, 500 for sentiment classification training, 50 for testing filter tweets, and 50 for testing sentiment classification. Methods TF-IDF produce the highest accuracy, with 74% on the classification filter tweets and 80% in klasfikasi sentiment, methods Ratio produces accuracy 72% on the classification filter tweets and 76% in klasfikasi sentiment, and N-Gram produce accuracy 64% on classification filter tweets and 78% in klasfikasi sentiment. So, for the case of sentiment analysis on twitter better use of feature selection methods TF-IDF that produces better accuracy in the classification stage.

3.2 Suggestion

Based on the results of testing that has been done, then the suggestions are expected after comparisons of feature selection for SVM sentiment classification on Twitter.

1. Use different kernels of SVM method for comparisons such as polynomial kernel, gaussian, or RBF.
2. Update a list of emoticons that can detect more emoticons including the latest emoticons.
3. Use more datasets.

BIBLIOGRAPHY

[1] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publisher, 2012.

[2] P. Koncz and J. Paralic, "An approach to feature selection for sentiment analysis," *IEEE International Conference*, p. 357–362, 2011.

[3] T. H. Susilo and S. Rochimah, "Pengklasifikasian Topik dan Analisis Sentimen Dalam Media Sosial," *SNASTI*, 2013.

[4] A. R. Sentiaji dan A. M. Bachtiar, "Analisis Sentimen Terhadap Acara Televisi Berdasarkan Opini Publik," *Jurnal Ilmiah Komputer dan Informatika*, 2014.

[5] P. Pitria, "Analisis Sentimen Pengguna Twitter pada Akun Resmi Samsung Indonesia dengan menggunakan Naive Bayes," Perpustakaan Unikom, 2014.

[6] F. Z Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," Universiteit van Amsterdam, Amsterdam.

[7] Octaviani. P. A., Wilandari. Y., & Ispriyanti. D., "Penerapan Metode Klasifikasi Support Vector Machine (SVM) Pada Data Akreditasi Sekolah Dasar (SD) Di Kabupaten Magelang," *Jurnal Gaussian*, vol. 3, pp. 811-820.

[8] I. Dilrukshi and K. d. Zoysa, "A Feature Selection Method for Twitter News Classification," *International Journal of Machine Learning and Computing*, vol. 4, no. 4, 2014.

[9] J. Fürnkranz, "A Study Using n -gram Features for Text Categorization," *Austrian Research Institute for Artificial Intelligence*, p. 8, 1998.

[10] D. S. Pamungkas, N. A. Setiyanto and E. Dolphina, "Analisis Pada Sosial Media Twitter Menggunakan Naive Bayes Classifier Terhadap Kata Kunci "Kurikulum 2013"," *Techno.COM*, vol. 14, no. 4, 2015.

[11] F. T. Anggraeny and M. Widyasri, "Klasifikasi Voting ANN PSO Biclass dengan Seleksi Fitur Gian Ratio," 2012.