SUPPORT VECTOR MACHINE AND INFORMATION GAIN FOR EMOTION CLASSIFICATION IN SONG LYRIC

Eva Nur Azizah1, Ednawati Rainarli2

[1.2]Indonesian Computer University
Jl. Dipatiukur No.112-116, Lebakgede, Coblong, Bandung, West Java 40132
E-mail: eva@email.unikom.ac.id[1], ednawati.rainarli@email.unikom.ac.id[2]

ABSTRACT

One form of text that can express the emotions are the lyrics of the song. Previous research conducted by Isra Citrawati Salekhah who studied song obtained an accuracy of only 36.66%. This is because the training data used English so that when translated into Indonesian goes wrong meaning of vocabulary (lexical) which resulted in an error meaning is conveyed. Therefore in this study will be used training data Indonesian songs that have been validated by linguists. Information Gain Selection feature will be added in order to obtain the relevant features in detecting emotions in song. The test results showed that the use of Support Vector Machine and Information Gain feature selection with training data of 1000 lines of lyrics and test data as much as 300 rows lyrics produce the best accuracy of 73.3%. These results were obtained by using a kernel functionpolynomialdegree 1. As for the results of tests on the emotions as much as 20 tracks show that the use of Support Vector Machine and Information Gain feature selection to get the best accuracy by 70%. This shows that the use of Information Gain feature selection can improve the accuracy for features that are not relevant to the classification of the target has been reduced.

Keywords :*Support Vector Machine, Information Gain,* classification, emotion, song

1. PRELIMINARY

Emotions can be expressed by a person of words, facial expressions, tone of voice and text [1]. One form of text that can express the emotions are the lyrics of the song. The lyrics are literary works in the form of a text whose contents can express the outpouring of personal feelings, thoughts and emotions of the songwriter. Therefore, the lyrics can be used as a research object in the classification of emotions.

In previous studies on the classification of the news carried by Siti Nur Assia showed that the method of Support Vector Machine has better performance than the K-NN method, where accuracy in Support Vector Machine reached 93.2% [2]. However, the research results Siti N different from those obtained by Isra Citrawati Salekhah who studied Indonesian song lyrics using the Multi-Class Support Vector Machine with TF-IDF weighting. The result shows that the accuracy of the study has been 36.66%. The test results of these studies indicate on training by using more data, accuracy tends to lower accuracy of the data is less. This is because the training data used English so that when translated into Indonesian goes wrong meaning of vocabulary (lexical) which resulted in an error meaning conveyed [3]. ThereforeThis research will be used in the training data Indonesian songs that have been validated by linguists, Information Gain Selection feature will be added in order to obtain the relevant features in detecting emotions in song,

*Support Vector Machine*a method that is widely used mainly in data classification process. Moreover, the SVM also has advantages in processing high-dimensional data without compromising performance [4]. It can be shown in a study conducted by comparing Joachims SVM to text Categorization by some other method that Bayes, Rochio, R4.5, and K-NN is known that SVM produces a good performance, outperformed the other methods are substantially and significantly [4]. The use of Support Vector Machine method showed better results in terms of accuracy and speed of manufacture of the model [12].

information Gain is one of the feature selection that can be used to select the best features and informative. Information Gain can see every feature to predict the correct class label for selecting the highest value and are important in order to improve the performance of classification algorithms [5]. This is shown in research conducted Nice Setya in text categorization by using Support Vector Machine that the addition of Information Gain feature selection may show an increase of 15% precision and recall by 13% compared with feature selection Particle Swarm Optimization (PSO) [6].

Therefore, in this study will be used method of Support Vector Machine and Information Gain feature selection to look at the accuracy of the method.

The purpose of this research is to implement feature selection classification Information Gain on

Support Vector Machine method for classification of emotions in the song lyrics. While the goal to be achieved in this research is to measure the performance of the algorithm feature selection algorithms Information Gain on Support Vector Machine classification method to determine the emotion in the song lyrics.

2. THEORETICAL BASIS

emotion

Emotion is a picture of a person caused by important events. Emotions include conscious mental state, physical disorder in some organs of the body, and facial expression recognition [7].

Type of basic emotions in humans there are five kinds:

a. Happy, describing someone who is successful or moving towards the success of a destination
b. Sad, illustrate the failure or loss
c. Angry, a person who was frustrated from the role or purpose other people feel
d. Fear, describes the physical or social threat to yourself

preprocessing

*Preprocessing*is is the stage to prepare the text into data to be processed in the next stage. Text to be done on this process be noise and there is not a good structure [8]. Preprocessing stages in this study is a case folding, convert negation, filtering, tokenizing and stopword removal.

a. *Case folding*
   *Case folding*a process that is done to unify the character data (documents / text). At this stage all homogenized to lowercase characters (lowercase) [8].
b. *Convert negation*
   *Convert negation*a conversion process of negation words contained in a sentence, for the word of negation has an influence in changing the value of emotion in a sentence [11]. Examples of negation words in Indonesian is "no", "no" and "not".
c. *tokenizing*
   *tokenizing*namely decomposition song in the form of a sentence - the sentence into words. Tokenizing process is done by separating the words by a space located between the two words [11].
d. *filtering*
   Filtering is the stage of selection of words - an important word, that word of what will represent the content of the document. At this stage characters other than letters removed and considered delimiter [11].
e. *stopword Removal*
   *stopword Removal*is a process to eliminate irrelevant words in the document text by comparing with existing stoplist. Stoplist contains a set of words that are not relevant but often appears in a document. In this study stoplist used was taken from Tala Z. Fadillah study [9].

TF-IDF Weighting

*Term Frequency - Inverse Document Frequency*used to determine the weight of a word in the many documents [16]. Weighting can be obtained based on the number of occurrences of a word (term) in a document collection term frequency (tf) and the number of occurrences of the term in a document collection inverse document frequency (idf). IDFT value of a term can be calculated using equation (1):

$$idf_t = \log(\frac{N}{df}) \quad (1)$$

Information :

N = Number of documents

df = Appearance of documents containing the terms

The equation used to calculate the weight (W) each - each document that is using equation (2):

$$W_t = tf_{dt} * idf_t \quad (2)$$

Information :

tf_{dt} =number of occurrences of term t to the document d

idf_t =the result of the inverse document frequency of the term t

information Gain

*information Gain*is one of the feature selection algorithm is used to select the best features. Gain Value Information obtained will be used to select features using a threshold so as to produce the best features [15]. Information Gain Value can be defined by equation (3) [10]:

$$IG(t) = -\frac{A+C}{N} log\left(\frac{A+C}{N}\right) + \frac{A}{N} log\left(\frac{A}{A+B}\right) + \frac{C}{N} log\left(\frac{C}{C+D}\right) \quad (3)$$

Information :

A = The number of documents that contain term k class t

B = Number of documents outside of class k containing term t

C = The number of documents that do not contain a class term k t

D = Number of documents outside of class k that does not contain the term t

N = Total document

After the Information Gain value obtained, then the next step is to determine the threshold value. Information Value Gain the smallest can be used as the threshold value with the best accuracy [13].

Support Vector Machine

*Support Vector Machine*is a supervised learning algorithm that is used for classification analysis. Support Vector Machine is a binary classifier that divides the data into two classes called hyperlane [14]. This Hyperlane right in the middle - the middle class. However Support Vector Machine has been developed in order to work in the case of non-linear using a kernel concept in high-dimensional

space. Kernel function is used to map the initial feature set lower on the set of new features higher. Sorts - kind of kernel functions can be seen in Table 1 [15]:

Table 1. Kernel functions

name Kernel	Function definition
linear	$K(x, x_k) = x.x_k^T$
polynomial	$K(x, x_k) = (x.x_k^T + 1)^d$
Gassian RBF	$K(x, x_k) = exp\left(\frac{-\|x - x_k\|^2}{2.\sigma^2}\right)$
sigmoid	$K(x, x_k) = \tanh[K\, x_k^T.x + \theta]$

Hyperlane SVM classification can be expressed by equation (4):

$$w.x_i + b = 0 \quad (4)$$

If the data is the data and the label is = -1 (negative), it can be expressed by equation (5) as follows:$x_i y_i$

$$w.x_i + b \leq -1 \quad (5)$$

Meanwhile, if the data is the data and the label is = +1 (positive) then can be expressed by equation (6):$x_i y_i$

$$w.x_i + b \geq +1 \quad (6)$$

To search for the best hyperlane can use the Quadratic Programming (QP) is to minimize the equation $\frac{1}{2}||w||^2$ using Lagrange Multiplier function which is defined as into the equation (7):

$$L_P = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \alpha_i\, y_i(w.x_i - b\sum_{i=1}^{N} \alpha_i y_i + \sum_{i=1}^{N} \alpha_i \quad (7)$$

Where is is lagrange multiplier berkorespodensi with. Value is zero or a positive value (≥ 0). The equation above will be turned into a form of duality Lagrange multiplier by maximizing the equation (8):$\alpha_i x_i \alpha_i \alpha_i L_D$

Maximize:

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i x_j} \quad (8)$$

With the proviso 1:

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad (9)$$

And condition 2:

$$\alpha_i \geq 0, i = 1,2, \ldots, N \quad (10)$$

With the dot-product of two data in the training data. Once the solution is found Quadratic Programming (value), then the class of test data can be determined by equation (11):$x_i x_j \alpha_i$

$$f(\Phi(x)) = sign(w.\Phi(z) + b) = sign\,(\sum_{i=1}^{N} \alpha_i y_i \Phi(x_i).\Phi(z) + b) \quad (11)$$

Where N is the number of data into a support vector, is a support vector, and z is the test data to be input.x_i

3. RESEARCH METHODS

Steps being taken in this study are in Figure 1 as follows:

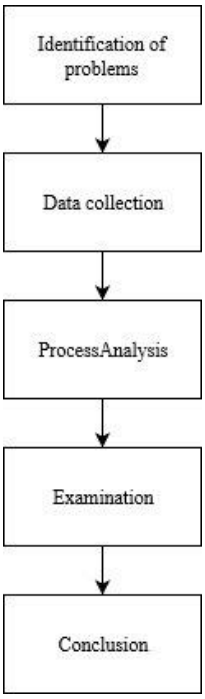

Figure 1. Stages of research used

Here is an explanation of each of the steps being taken in this study:

a. Identification of problems

At the stage of identification of problem analysis on the problems faced in the research undertaken.

b. Data collection

At the stage of data collection that collects data and data song literature. Required literature is the method of Support Vector Machine and Information Gain. While the lyrics of data taken from various sites https://liriklaguindonesia.net/, https: //no1lyrics.com/, and other sites. Each line of the lyrics will be validated by linguists based on the type of emotion.

c. Process Analysis

In the analysis phase analysis process regarding the classification process emotions. Analysis was conducted on the analysis of the problem, analysis preprocessing, feature selection analysis and analytical methods Information Gain Support Vector Machine.

d. Examination

At this stage will be tested against the classification results by using Conffusion Matrix.

e. Conclusion

At this stage is to determine the conclusions based on the results of testing that has been done.

4. RESULTS AND DISCUSSION

Problem analysis

Problems found in previous research is that training data used to speak English so that when translated into Indonesian goes wrong meaning of vocabulary. The solution provided is to replace the training data with Indonesian lyrics that have been validated by

linguists and adding Gain Information feature selection for feature selection can obtain the relevant features in detecting emotions in the song so that it can improve the performance of classification algorithms.

Input Data Analysis

Data will be divided into two types of training data and test data. Training data used is a collection of sentences as many as 1000 lines of lyrics that have been grouped by the type of emotion that is determined. While the test data used is a collection of sentences of 300 lines of lyrics. In this study, there are four categories of emotions that have been determined are happy, sad, angry and afraid.

Process Analysis

Stages are used to determine what kind of emotion in the lyrics in this study were divided into two phases: a training phase and a testing phase.

in stagetraining training data will go through four main processes preprocessing process, TF-IDF weighting, Information Gain feature selection and training process Support Vector Machine. In the testing phase, the test data is entered through the same process as training data is the process of preprocessing, TF-IDF weighting, and the testing process with Support Vector Machine. The results of this testing phase is the classification results in the form of emotional categories of test data that has been entered. Then for system testing are presented in confuss matrix. Results from these tests is the comparison of the accuracy of the method Support Vector Machine without feature selection Gain Information and Support Vector Machine with Information Gain feature selection.

Testing Accuracy

Testing accuracy is divided into three sections, where each section will use a varied amount of test data. In the first test is to determine the accuracy of the classification model, test data are used together with training data used is 1000 lines of lyrics. In the second test used 1000 training data and test data lyric line 300 line lyrics. In the third test, namely the test of emotions song by testing as many as 20 tracks and training data of 1000 lines of lyrics.

Results of testing the accuracy of results of the test of the accuracy menggunakann method of Support Vector Machine and Information Gain. The following test results:

a. Classification Model Testing Results

Here is a model of the work of the classification by using the same training data with test data with the data number 1000 lines of lyrics, accuracy results can be seen in Table 2:

Table 2. Classification Model Testing Results

	SVM	**SVM + IG Threshold = 0.00181**	**SVM + IG Threshold = 0.00183**
linear	88.8%	86.7%	80,4%
RBF γ =1	94.6%	93.2%	84.2%
RBF γ =2	96.2%	95.3%	85.1%
RBF γ =3	96.2%	95.3%	85.1%
Polynom n = 1	94.5%	93.4%	85%
Polynom n = 2	96.2%	95.3%	85.1%
Polynom n = 3	96%	95.3%	85.1%

The results of these tests menujunkan that the best model is generated by Support Vector Machine with RBF kernel function andpolynomialproduces an accuracy of 96.2%. This shows that the use of Information Gain feature selection is not too influential to the same data.

b. The assay results with different test data

Here is the result of classification using training data of 1000 lines of lyrics and test data as much as 300 lines of lyrics, which can diihat in Table 3:

Table 3, Testing Results with 300 lines of lyrics

	SVM	**SVM + I Threshold = 0.00181**	**SVM + IG Threshold = 0.00183**
linear	64.6%	66.6%	63.3%
RBFγ =1	66%	68%	65%
RBFγ =2	67.6%	70%	65,3%
RBFγ =3	65%	67%	66.3%
Polynom n = 1	70.3%	73.3%	65%
Polynom n = 2	65%	66.3%	62%
Polynom n = 3	55,6%	61.3%	62%

The test results are directing that the best method is generated by Support Vector Machine with first degree Polynom kernel function using Information Gain feature selection. The results obtained accuracy of 73.3%. This shows that the use of Information Gain feature selection can improve the accuracy for

features that are not relevant to the classification of the target has been reduced.

c. Testing Emotional Songs

In this test 2 is divided into testing using *Support Vector Machine*Gain without feature selection Information and Support Vector Machine with Information Gain feature selection. Kernel function used is the kernel function polynom degree 1.

Results of testing the confusion matrix for testing tracks without feature selection can be diihat in Table 4:

Table 4. Testing confusion matrix Songs without feature selection

F_{ij}		Class prediction (j)			
		class 1	**Grade 2**	**Grade 3**	**4th grade**
The original class (i)	**class 1**	5	0	3	0
	Grade 2	3	2	0	1
	Grade 3	0	1	0	0
	4th grade	1	0	0	4

The test results show that the number of songs that were classified correctly as many as 11 songs and the wrong result as much as 9 songs. The results showed that the total songs in class 1 (excited) by 8 tracks with 5 songs of predicted correctly in class happy and 3 songs predictable in class 3 (angry).

Then the value calculation accuracy by looking at the confusion matrix obtained by 55%.

$$Accuracy = \frac{11}{20}\ x\ 100\% = 55\%$$

While the test results to the test track confusion matrix using feature selection can be diihat in Table 5:

Table 5. Testing confusion matrix Songs with feature selection

F_{ij}		Class prediction (j)			
		class 1	**Grade 2**	**Grade 3**	**4th grade**
The original class (i)	**class 1**	8	0	0	0
	Grade 2	3	2	0	1
	Grade 3	0	0	1	0
	4th grade	2	0	0	3

The test results show that the number of songs that were classified correctly as many as 14 tracks and the wrong result as 6 songs. The results showed that the total songs in class 2 (sadly) as 6 songs with 2 tracks of predicted correctly in class sad, happy class 3 predictable song and one song in grade 4 (fear).

Then the value calculation accuracy by looking at the confusion matrix obtained by 70%.

$$Accuracy = \frac{14}{20}\ x\ 100\% = 70\%$$

5. CONCLUSION

The test results showed that the use of Support Vector Machine and Information Gain feature selection with training data of 1000 lines of lyrics and test data as much as 300 rows lyrics produce the best accuracy of 73.3%. These results were obtained by using a kernel functionpolynomialdegree 1. As for the results of tests on the emotions as much as 20 tracks show that the use of Support Vector Machine and Information Gain feature selection to get the best accuracy by 70%. This shows that the use of Information Gain feature selection can improve the accuracy for features that are not relevant to the classification of the target has been reduced. Suggestions for kedepanya can apply deep learning for classification.

BIBLIOGRAPHY

[1] V. V. Ramalingam, A. Pandian, A. Jaiswal, and N. Bhatia, "Emotion detection from text," J. Phys. Conf. Ser., vol. 1000, no. 1, 2018.

[2] S. N. Asiyah and K. Fithriasari, "Klasifikasi Berita Online Menggunakan Metode Support Vector Machine dan K- Nearest Neighbor," *J. Sains dan Seni ITS*, vol. 5, no. 2, pp. 317–322, 2016.

[3] R. Baharuddin, "Kesalahan Makna Leksikal pada Terjemahan Teks Bahasa Indonesia ke dalam Bahasa Inggris," Dialekt. J. Lang. Lit. Math. Educ., vol. 1, no. 1, pp. 42–55, 2015.

[4] I. K. Purnamawan, "Support Vector Machine Pada Information Retrieval," Jptk, vol. 12, no. 2, pp. 173–180, 2015.

[5] I. Maulida, A. Suyatno, and H. R. Hatta, "Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain," Jsm Stmik Mikroskil, vol. 17, no. 2, pp. 249–258, 2016.

[6] B. S. Rintyarna and A. Z. Arifin, "Seleksi Fitur Dua Tahap Menggunakan Information Gain dan Artificial Bee Colony untuk Kategorisasi Teks Berbasis Support Vector Machine," Systemic, vol. 1, no. 2, pp. 22–26, 2015.

[7] S. Sumpeno, "Klasifikasi Emosi Untuk Teks Bahasa Indonesia," Semin. Nas. Pascasarj. IX – ITS, 2009.

[8] I. Feinerer, "Mining Text Data," R News, vol. 8, pp. 51–88, 2012.

[9] F. Z. Tala, "A Study Of Stemming Effects On Information Retrieval In Bahasa Indonesia," M.Sc. Thesis, Append. D, Vol. Pp, Pp. 39–46, 2003.

[10] M. Fatih And S. Bayir, "Examining The

Impact Of Feature Selection Methods On Text Classification," IJACSA) Int. J. Adv. Comput. Sci. Appl., Vol. 8, No. 12, Pp. 380–388, 2017

[11] R. Fahreza Nur Firmansyah, M. Fauzi, T. Afirianto, "Sentiment Analysis Pada Review Aplikasi Mobile Menggunakan Metode Naïve Bayes Dan Query Expansion, " Vol. 8, 2016.

[12] E. Rainarli and A. Romadhan, "Perbandingan Simple Logistic Classifier dengan Support Vector Machine dalam Memprediksi Kemenangan Atlet," J. Inf. Syst. Eng. Bus. Intell., vol. 3, no. 2, pp. 87–91, 2017.

[13] R. M. Alfajri, Y. H. Chrisnanto, And R. Yuniarti, "Pengklasifikasian Kemampuan Akademik Mahasiswa Menggunakan Metode Information Gain Dan Naive Bayes Classifier Dalam Prediksi Penyelesaian Studi Tepat Waktu," Pros. Snst, No. 7, pp. 144–149, 2016.

[14] E. Prasetyo, Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab, Andi, 2014.

[15] Suyanto, Data Mining Untuk Klasifikasi Dan Klasterisasi Data, Informatika, 2017.

[16] D. S. Harjanto, S. N. Endah, and N. Bahtiar, "Sistem Temu Kembali Informasi pada Dokumen Teks Menggunakan Metode Term Frequency Inverse Document Frequency (TF-IDF)," J. Sains dan Mat., vol. 20, no. 3, pp. 64–70, 2012.