DOCUMENT CLASSIFICATION WITH SUPPORT VECTOR MACHINE AND MUTUAL INFORMATION

Sherly Amanda Kristiani[1] , Ednawati Rainarli[2]

[1,2] Informatic Engineering – Indonesian Computer University
Jalan Dipatiukur 112-116 Bandung
E-mail : sherlyamanda007@gmail.com[1], ednawati.rainarli@email.unikom.ac.id[2]

ABSTRACT

Documents classification is the process of grouping documents into one category which features the same word. This study will classify the document abstract thesis is based on machine learning using Support Vector Machine (SVM) and feature selection in the form of the Mutual Information (MI). Selection feature Mutual Information (MI) is used to determine a word that is characteristic or unique words used in the document abstract thesis that would be classified. The goal for abstract document classification using Support Vector Machine (SVM) may result in better accuracy. Steps being taken include preprocessing, tf-idf weighting, feature selection Mutual Information (MI) and the classification of documents to determine the category of abstract thesis. Testing is done by comparing the classification with Mutual Information (MI) feature selection process and without feature selection. Results of testing for the use of Mutual Information (MI) obtained an accuracy of 94%, and without the use of Mutual Information (MI) by 94%. Based on these results concluded that the use of Mutual Information (MI) feature selection on an abstract document classification does not have significant differences.

Keywords: Documents Classification, SVM, Mutual Information (MI).

1. PRELIMINARY

Document is a collection of writings that contain information, printed or electronic [1]. When someone searches for information it is very easy to miss unexpected categories or in the classification process of the document.

Previous research conducted by Ahmad Yusuf explained the classification of documents with SVM and K-Means Clustering to increase the accuracy value of 88.1% from 5 groups. The SVM method has good performance supported by K-Means Clustering [2]. While research by Amanda Nurul Amalia, explains the Implementation of SVM in the Thesis Report Classification produces an accuracy of 34%. The results of research testing tend to be small, because the word features used in the word weighting process are not relevant in determining the class to be classified in the classification process [3].

To be able to produce greater accuracy in the abstract document classification process, Mutual Information (MI) and Support Vector Machine (SVM) are used. Julius Gigih Dimastyo's study explains the comparison of spam filter feature selection in MNB classification cases assisted by the selection of Idf, Mutual Information (MI) and Chi Square features in improving classification accuracy. From the research produced a mutual information method with an accuracy of 93.77% [4].

Because of the low classification process of abstract documents with SVM accuracy is owned. So this research aims to improve accuracy with the Support Vector Machine method and Mutual Information feature selection in the case of document classification.

The amount of data is 200 abstracts where 150 abstracts become training data and 50 abstracts become test data. The training data and test data itself will be in the form of .csv. The abstract document classification process aims to group abstract documents into 5 scientific group categories contained in the Informatics Engineering Study Program, where each group has unique word features so that it can help the document classification process to be carried out. The abstract data used includes titles, abstracts and their contents, divided into scientific groups based on their unique characteristics.

2. LITERATURE REVIEW

Document classification is the process of grouping documents [5]. Document classification giving a category, to documents that do not yet have a category [6]. Thus, someone looking for information is easier to miss categories that are not the purpose.

2.1 Preprocessing

Preprocessing is the sequence of preparation of text into processed data in the next process. Initial input in the form of abstract documents.

2.2.1 Case Folding

Case Folding is the act of changing all the letters in a document to lowercase [7] [8].

2.2.2 Filtering

Filtering is done by taking important words from the token results [7].

2.2.3 Tokenizing

Tokenizing is the separation of every word contained in a document by removing punctuation, characters and numbers other than the alphabet [7] [8].

2.2.4 Stopword Removal

Stopword Removal is the act of deleting words contained in a stoplist. Stoplist contains a collection of words that are not relevant but often appear in a document [7] [8].

2.2 TF-Idf Word Weighting

Word weighting is the process of determining word frequency values in documents [9]. The idf (word) value is calculated by the following equation (1):

$$IDF = \log(N/df) \tag{1}$$

Weight value (W) calculated by equation (2) :

$$W = tf \cdot IDF \tag{2}$$

2.3 Mutual Information (MI)

Mutual Information (MI) is a feature selection method for classifying a document, so that it is efficient and effective by reducing features [8]. MI can also be used to measure the number of attributes that play a role in making a correct classification, so that it will produce input that is more influential on the classification process [10]. The Mutual Information (MI) value can be calculated by equation (3):

$$\mathrm{MI}(t,c) \approx \log \frac{A*N}{(A+C)*(A+B)} \tag{3}$$

2.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a method used to find hyperplane so that it can separate two data sets in two different classes. Hyperplane is a data dividing line between classes [6] :

$$L_D = \sum_{i=1}^{n} \alpha i - \frac{1}{2} \sum_{i=1}^{n} \alpha_i \alpha_j \, y_i y_j \, x_i x_j \tag{4}$$

Condition 1 :

$$\sum_{i=1}^{n} \alpha_i y_j = 0 \tag{5}$$

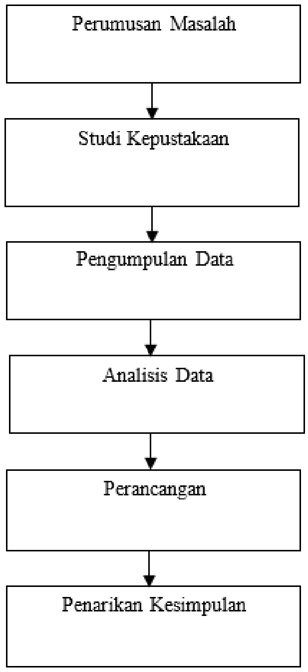
Condition 2 :

$$\alpha_i \geq 0, i = 1,2, \dots, N \tag{6}$$

N is the amount of vector support data, x_i is a support vector, z is a test data that the class will predict, and $x_i \cdot z$ is an inner-product intermediate xi and z.

3. RESEARCH METHODS

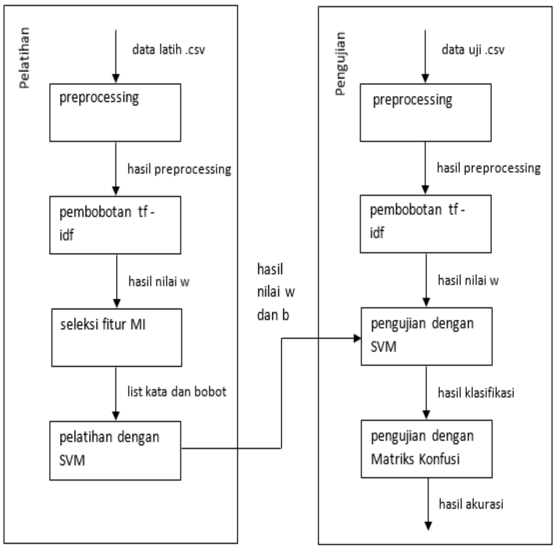
The research method is the steps used when conducting research, what will be done, how the process is going through, how to analyze. In this study through several stages carried out as in Picture 1 according to S. Guritno [11]. The steps contained in the research method, including:



Picture 1. Research Methods

3.1 System Analys

Document classification process has several stages of analysis. The steps carried out in the System Analysis according to Picture 2 are as follows:



Picture 2. Document Classification System Analysis Performed

3.1 Process Analysis

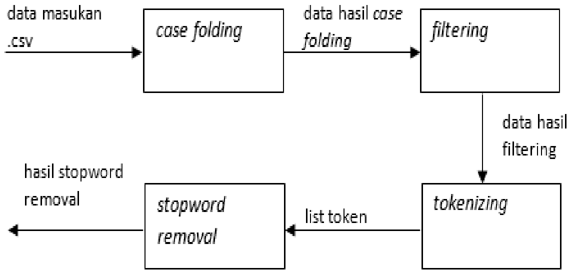
The steps taken in the abstract document classification process:

1. Prepare input data to be processed as an example in table 1

Table 1. Sample Input Data

Abstrak	Kategori
SISTEM INFORMASI MANAJEMEN PROYEK DI CV. MUKTI JAYA ABSTRAK CV. Mukti Jaya merupakan salah satu perusahaan yang bergerak dalam bidang jasa konstruksi. Berdasarkan permasalahan yang ada, maka dibutuhkan sebuah Sistem Informasi Manajemen Proyek untuk menangani jadwal perencanaan proyek, pengawasan biaya dan waktu, dan pengelolaan risiko. Metode Critical Path Method (CPM) digunakan untuk perencanaan jadwal. Metode Earned Value Management (EVM) digunakan untuk pengawasan biaya dan waktu proyek. Sedangkan untuk mengelola risiko menggunakan metode Probability Impact Matrix (PIM) dan metode Earn Value Management (EMV). Kata kunci : Manajemen Proyek, Sistem Informasi, Critical Path Method, Earned Value Management, Probability Impact Matrix, Expected Monetary Value.	A

2. The preprocessing process is a stage where the abstract is prepared into data to be processed to the next stage. Preprocessing steps carried out are in Picture 3 as follows:



Picture 3. Preprocessing

After doing the preprocessing stage, the results of the stopword removal process are in table 2.

Table 2. Preprocessing Result

Preprocessing Result		
sistem	informasi	manajemen
mukti	jaya	abstrak
jaya	salah	perusahaan
jasa	konstruksi	berdasarkan
sistem	informasi	manajemen
jadwal	perencanaan	proyek
pengelolaan	risiko	metode
method	cpm	perencanaan
earned	value	management
biaya	proyek	mengelola
probability	impact	matrix
earn	value	management
manajemen	proyek	sistem
path	method	earned
probability	impact	matrix
proyek	evm	cv
cv	risiko	mukti
bergerak	pim	bidang
permasalahan	emv	dibutuhkan
proyek	informasi	menangani
pengawasan	value	biaya
critical	expected	path
jadwal	evm	metode
pengawasan	kunci	monetary
metode	critical	value
metode	management	

3. Calculate the weights to get the IDF value and produce 147 words.
   For example, for system words contained in D1 and D5, the IDF and W values will be calculated as follows:

Diketahui : tf = 3 (pada D1),
tf = 2(pada D5),
df = 2, N = 5

$IDF_{(sistem)}$ = $\log(N/df)$
= $\log(5/2)$
= 0.3979

$W_{(sistem)}$ = $tf \cdot IDF$
= 3 * 0.3979
= 1.1937 (Untuk $_{(sistem)}$ pada D1)

$W_{(sistem)}$ = $tf \cdot IDF$
= 2 * 0.3979
= 0.7958 (Untuk $_{(sistem)}$ pada D5)

4. Select the MI feature for each category, get a feature that has a Mutual Information value of 0.6989 for the next process. Example of calculating the MI value of each word in D1 - D5 categorized A - E, for example for words:

$$I_{(sistem)} = \log \frac{A.N}{(A+B)(A+C)} = \log \frac{1.5}{(1+1)(1+0)} = \log [5/2] = 0.3979$$

$$I_{(visualisasi)} = \log \frac{A.N}{(A+B)(A+C)} = \log \frac{1.5}{(1+0)(1+0)} = \log [5/1] = 0.6989$$

5. Perform SVM calculations for training.
In the Lagrange multiplier duality equation we get:
Maximize:

$$L_D = \sum_{i=1}^{n} \alpha i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, y_i y_j \, x_i x_j$$

$$L_D max = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 - \frac{1}{2} (57.14 \, \alpha_1^2 + 14.16 \, \alpha_2^2 + 1.95 \alpha_2 \alpha_3 + 1.95 \alpha_3 \alpha_2 + 33.7 \, \alpha_3^2 + 37.12 \, \alpha_4^2 + 32.23 \, \alpha_5^2)$$

Condition 1 : $\alpha_1 - \alpha_2 - \alpha_3 - \alpha_4 - \alpha_5 = 0$

Condition 2 : $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5 \geq 0$

In the objective function, the second term is multiplied by $y_i y_j$. The equation meets the standard Quadratic Programming so that it can be helped to solve it with a commercial solver for Quadratic Programming (QP). With the help of the software, the following results were obtained:

$\alpha_1 = 0.031$
$\alpha_2 = 0.013$
$\alpha_3 = 0.005$
$\alpha_4 = 0.005$
$\alpha_5 = 0.006$
b = -101.41

These results indicate that all training data are support vectors due to the value of a > 0. While the value of b is obtained from the training process carried out. After finding all a and b, the SVM model can be used for the prediction model by using:

$$f(x) = w^T . x + b$$

$$(x) = a^T \, \bar{y} \, (x_{i,uji}) + b$$
$$f(x) = 0.031 \, . \, (1) \, . \, K \, (x_{1,uji}) - 0.013 \, . \, (-1) \, . \, K \, (x_{2,uji}) \ldots - 0.006 \, . \, (-1) \, . \, K \, (x_{5,uji}) - 101.42$$

With xi is a support vector and $x$test is a test data to be predicted.

6. Perform SVM calculations for training.

$$kelas \ x = \arg \max_{k=1,\ldots,5} \begin{pmatrix} [w^1]^T . \varphi(x) + b^1, \\ [w^2]^T . \varphi(x) + b^2, \ldots, \\ [w^5]^T . \varphi(x) + b^5 \end{pmatrix},$$

The values of w1 and b1 were obtained from the results of training that had been carried out previously where number 1 shows the index of class 1, namely category A, number 2 shows index of class 2, namely category B, number 3 indicates index of class 3, namely category C, number 4 indicates index of class 4 namely category D and number 5 indicate the class 5 index namely category E. Class calculations are based on predictive models:

$$kelas \ x = \arg \max_{k=1} = ((0.031 * 36.95) - (0.013 * 0) - (0.005 * 0) - (0.005 * 0) - (0.006 * 0)) - 101.41 = -100.27$$

$$kelas \ x = \arg \max_{k=2} = ((-0.014 * 36.95) + (0.08 * 0) - (0.019 * 0) - (0.021 * 0) - (0.025 * 0)) + 45.51 = 45.05$$

$$kelas \ x = \arg \max_{k=3} = ((-0.006 * 36.95) - (0.019 * 0) + (0.047 * 0) - (0.009 * 0) - (0.011 * 0)) + 20.05 = 19.83$$

$$kelas \ x = \arg \max_{k=4} = ((-0.005 * 36.95) - (0.021 * 0) - (0.008 * 0) + (0.045 * 0) - (0.009 * 0)) + 17.29 = 17.11$$

$$kelas \ x = \arg \max_{k=5} = ((-0.006 * 36.95) - (0.024 * 0) - (0.009 * 0) - (0.009 * 0) + (0.05 * 0)) + 20.06 = 19.84$$

In order to get the value of each class:

$$kelas\ x = \arg\max_{k=1\ldots5} (-100.27,\ 45.05,\ 19.83,\ 17.11,\ 19.84)$$

$kelas\ x = 45.05$

The biggest hyperplane value is 45.05 where the hyperplane value is a class 2 hyperplane value. It means that the P_Uji data is an abstract document with category B.

4. RESULT AND DISCUSSION

Accuracy testing is a stage that has the objective to find out the level of accuracy of using Support Vector Machine and Mutual Information by calculating the amount of test data whose class is predicted correctly. How to measure the performance of the system will be done using a confusion matrix. Confusion matrix is a table that records the results of classification work. The training data used were 150 data, while the test data used were abstract documents of 50 data.

a. Testing Results of Support Vector Machine Models use 150 Training Data and Test Data.
   The training data is the same as the test data with a total of 150 abstract documents, the results of the accuracy can be seen in table 3 and table 4.

Table 3. Test Results of Linear Classification Model and RBF

Condition	Linear	RBF		
		γ=1	γ=2	γ=3
SVM dengan Mutual Information (MI), Nilai MI sebesar = 0.6989	95%	96.33 %	96.34 %	98%
SVM tanpa MI	96%	97.33 %	97.34 %	98%

Table 4. Test Results of Linear Classification Model and RBF

Condition	Linear	Polynom		
		n=1	n=2	n=3
SVM dengan Mutual Information (MI), Nilai MI sebesar = 0.6989	95%	96.33 %	97.66 %	97.66 %
SVM tanpa MI	96%	97.33 %	98.67 %	99%

b. Vector Machine Support Model Testing Results use 150 Training Data and 50 Exam Data the results of its accuracy can be seen in tables 5 and table 6.

Table 5. Test Results of Linear Classification Model and RBF

Condition	Linear	RBF		
		γ=1	γ=2	γ=3
SVM dengan Mutual Information (MI), Nilai MI sebesar = 0.6989	94%	89%	86.34 %	81%
SVM tanpa MI	94%	90%	86%	82%

Table 6. Test Results of Linear and Polynom Classification Models

Condition	Linear	Polynom		
		n=1	n=2	n=3
SVM dengan Mutual Information (MI), Nilai MI sebesar = 0.6989	94%	80.33 %	80.33 %	80.33 %
SVM tanpa MI	94%	80%	80%	80%

Based on the results of tests conducted, the test results show that the use of Support Vector Machine and Mutual Information feature selection with 150 training data and 50 document data test results in an accuracy of 94%.

5. CONCLUSION AND SUGGESTION

Based on the results of research conducted in the Classification of Documents Using the Support Vector Machine and Mutual Information, the authors can draw some conclusions:

1. The highest accuracy value of the SVM method by using MI for abstract document classification is 94%. The addition of training data causes the Support Vector Machine method to increase accuracy because it contains many features that are obtained to help the classification process itself while for the use of Mutual Information feature selection itself does not affect too much, tends to get the same accuracy results and almost close
2. The training process carried out on 5 document categories obtained 147 words of weighting results and 137 words of feature selection results. This means only 10 words that are not used to carry out the training and testing process.

While suggestions for research that have been carried out in the Document Classification Using Support Vector Machines and Mutual Information are:

1. Based on the results of research that has been done, it still needs to be done several further studies. As for suggestions for further research, it is necessary to use various types of abstract documents with the addition of training data and test data will be more visible difference.
2. In addition, in the selection of training data and test data in abstract documents must be careful because not every document can be processed to make a category there are several factors that cause the document can not be defined.

.

BIBLIOGRAPHY

[1] A. Munandar, A. Hidayatno and T. Prakoso, "Klasifikasi Citra Dokumen Menggunakan Metode Support Vector Machine Dengan Ekstraksi Ciri Term Frequency - Inverse Document Frequency," *Transient,* vol. 6, no. 4, pp. 622-628, 2017.

[2] A. Yusuf and T. Priambadha, "Support Vector Machine Yang Didukung K-Means Clustering Dalam Klasifikasi Dokumen," *JUTI,* vol. 11, no. 1, pp. 13-16, 2013.

[3] A. N. Amalia, "Implementasi Support) Vector Machine (SVM) Pada Klasifikasi Laporan Skripsi (Studi Kasus Teknik Informatika)," Unikom, Bandung, 2016.

[4] J. G. Dimastyo, "Perbandingan Seleksi Fitur pada Spam Filter Menggunakan Klasifikasi Multinomial Naive Bayes," Fakultas Matematika dan Ilmu Alam IPB, Bogor, 2014.

[5] M. F. Sianturi, A. and S. A. Faraby, "Klasifikasi Dokumen Menggunakan Kombinasi Algoritma Principal Component Analysis dan SVM," *e-Preceeding of Engineering,* vol. 4, no. 3, pp. 5140-5144, 2017.

[6] N. Indriani, E. Rainarli and K. E. Dewi, "Peringkasan dan Support Vector Machine pada Klasifikasi Dokumen," *Jurnal Infotel,* vol. 9, no. 4, pp. 416-421, 2017.

[7] N. H. Ayuning Sari, M. A. Fauzi and P. P. Adikara, "Klasifikasi Dokumen Sambat Online Menggunakan Metode K-Nearest Neigbor dan Features Selection Berbasis Categorical Proportional Difference," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 2, no. 8, pp. 2449-2454, 2018.

[8] F. S. Nurfikri, M. S. Mubarok and A. , "Klasifikasi Topik Berita Menggunakan Mutual Information dan Bayesian Network," *e-Proceeding of Engineering,* vol. 5, no. 1, pp. 1579-1588, 2018.

[9] J. P. and J. S. , "Implemetasi Maximum Marginal Relevence dan Matriks Cosine Similarity Pada Aplikasi Peringkasan Dokumen," in *Teknik Informatika Universitas Mataram*, Mataram, 2015.

[10] W. Fonda, "Pembentukan Daftar Kata Kunci Untuk Pengklasifikasian Opini Pada Media Sosial Dengan Pendekatan Korpus dan Kamus," Teknik Elektro dan Informatika ITB, Bandung, 2014.

[11] S. Guritno, Theory and Application of IT Research, Yogyakarta: Andi Publisher, 2011.