Technical co-Sponsor :

ISBN :
978-1-5386-3085-3
IEEE CATALOG NUMBER :
CFP17CUE-ART

iSemantic

2017 International Seminar on
Application for Technology of Information
and Communication ( iSemantic )

# PROCEEDINGS

## Empowering Technology for a Better Human Life

Email:
isemantic@lppm.dinus.ac.id
Website:
http://isemantic.dinus.ac.id/

Semarang | October 7th - 8th, 2017

2017 International Seminar
on Application for Technology of Information and Communication (iSemantic)
ISBN : 978-1-5386-3085-3
IEEE CATALOG NUMBER : CFP17CUE-ART

Organized by

LPPMUDINUS

**Lembaga Penelitian dan Pengabdian Masyarakat** Universitas Dian Nuswantoro Semarang

**The Faculty of Engineering** Universitas Dian Nuswantoro Semarang

# PROCEEDINGS

2017 International Seminar on Application for Technology of Information and Communication (iSemantic)

**Empowering Technology for a Better Human Life**

October 7th – 8th, 2017
Universitas Dian Nuswantoro
Semarang, Indonesia

ISBN: 978-1-5386-3085-3
IEEE Catalog Number: CFP17CUE-ART

# COPYRIGHT

**2017 International Seminar on Application for Technology of Information and Communication (iSemantic)**



IEEE Catalog Number: CFP17CUE-ART

ISBN: 978-1-5386-3085-3

# TABLE OF CONTENT

# Performance Improvement Of Support Vector Machine (SVM) With Information Gain On Categorization Of Indonesian News Documents

**Adhy Rizaldy[1], Heru Agus Santoso[2]**
[1,2]*Researcher at*
*Computer Science Faculty of Dian Nuswantoro University*
*Jln. Imam Bonjol No.207, Semarang*
Correspond: adhyc4n@gmail.com, heru.agus.santoso@dsn.udinus.ac.id

***Abstraction***

**More news articles which are unstoppable increasing, causing problems with grouping news according to appropriate kind of label. Therefore it is necessary to deal with the problem of grouping news by it's category like business news, political news, and sports news. The categorization of news document belong to text classification domain, a Machine Learning topic as an approach that addressed this problem. Various algorithms have been used in previous studies such as Bayesian techniques, k-Nearest Neighborhood, Neural Networks, and Support Vector Machine (SVM). This study provides an understanding of the SVM method for news categorization on Indonesian news dataset that contain several types of news category. Problems in text classification is the number of features that affecting classification performance with SVM. Use of Information Gain as feature selection improve accuracy than without any feature selection. Our model give satisfying result with 98,057 % accuracy of Indonesia news classification. Improvement 2,9 points from 95,11% by SVM technique without feature selection.**

**Keywords**: news document classification, text categorization, SVM classification, Information Gain, Indonesia news classification

## I. INTRODUCTION

Online news as a main source of daily information in our country, have been increasing significantly. Some of news portal like *kompas.com* and *detik.com* become favorite and accessed by citizens continuously. This made they added subdomains as some new category to spread different kind of readers. Some of that like *hot.detik.com*, *kompasiana* of kompas, *wolipop* of detik.com and many others. These categorization is somehow can be wrong overtime caused by the hugely articles published.

We found some of these, for example entertainment news of *kompas.com*. Many articles on 2008 about economy and government figures had placed in 'entertainment.kompas.com' domain. Another case in *detik.com*, some news about '*olahraga'* topic on 2010 had placed in '*nasional'* category. One factor of this problem could be human error. In order to minimize this problem, the media stakeholder need technique to manage the archived of news files well. Some research has done in text classification named document classification to occupy this.

Document classification problem for Indonesia news have used many approach. Jaafar and partners classified Indonesia and Malaysia news from two webportals of each country with kNN based on technique [1]. Even though this Neighborhood algorithm, suffered on performance when training data is quite big. But when not big enough isn't going to be optimal [1]. Asy'arie and partner used Naive Bayes (NB) to 250 Indonesia news articles [2]. NB kind of sensitive to the amount of training document data and had drawn of performance problem [2].

Some papers have implemented Support Vector Machine as main classifier. Lilliana and partner [3] did classification with SVM on 180 news articles of kompas.com. Their method produced 85% average accuracy. Document classification problem for Indonesia news used SVM algorithm in Khodra research [4]. They could handle 10.404 articles for categorization with satisfied results. They compared several techniques for automatic classification with multilabel based on approach. By SVMs as the binary classifier, this research had 78% of F-measure in result. From these papers [3, 4], the amount of dataset used is quite different, but decreased in result, although both generated by SVMs classifier.

From this gap we could conclude that general problem of text classification is big dimension of data. In other words, a lot of features conducted from text dataset can cut down the result. In order to fix that problem, implementation of SVM technique for automatic text classification needs dimension data reduction.

In this work we discuss about feature selection as one way of dimensional reduction of massive data to improve machine learning method. Some feature selection have implemented. Information Gain had used in [5][6] with significantly

improvement in result compared by previous research. In Vinita comparison, IG is better than forward and backward selection [7]. The aim of this paper is to prove that IG as feature selection in SVMs based document classification could become solution for the big amount of dataset problem.

This research uses a combination of popular Machine Learning (ML) algorithm, SVM as multiple classes news document based on classification and Information Gain as feature selection. We propose this model as improvement from previous researches. Next sections of this paper are constructed as follows. Section 2 presented the related works of Indonesia news document classification. Section 3 & 4 discuss our proposed model to enhance text categorization, experiments results, and several points discovered. Finally, conclusion and future work are given in section 5.

## II. RELATED WORK

The Indonesian news classification in Wirabuana's study uses hybrid method [8]. The first step is to calculate the term weights using TF-IDF, to know the frequency of occurrence of terms in each document. Second, group the sample data of each category with K-Means algorithm. Then the training data sample is used in the classification using k-Nearest Neighbors. The use of k-Means clustering is to deal with complexity issues in the k-NN classification. Using 500 data from detik.com and kompas.com, the results showed an accuracy of 87% with an average F-Measure evaluation value of 0.802 [8].

Ariadi and colleagues [9], in their publication compared the Naive Bayesian Classification (NBC) and Support Vector Machine methods for the Indonesia News classification. SVM is known to be very good on data with large dimensions. Through this paper authors show that SVM with linear kernel and RBF kernel produce same rate of classification accuracy. For time measurement SVM based side is much faster to get results compared with NBC way.

With the concept of multilabel classification, Rahmawati and Khodra [4] implements several label-classification algorithms, feature selection algorithms and two method weighting terms for two frequently used approaches: problem transformation and adaptation methods. Using the 10,000 Indonesian news articles, the author presented the combination of the TF-IDF feature weighting method, Symmetrical Uncertainty feature selection, Calibrated Label Ranking - which is the problem transformation and SVM approach as its sole labeling algorithm. From this combination, however the emergence of Out Of Vocabulary problems can be a topic of discussion for another subsequent study.

In Wibawa and Purwarianti's study, statistically based NERC has been performed on Indonesian news documents [10]. Wibawa and Purwianti compare some single machine learning algorithms as well as an ensembled technique. Using 457 news articles, the best accuracy is achieved by using an ensemble technique in which the results of some machine learning algorithms are used as a feature for a single machine learning algorithm. Of the 457 news articles included in several categories, the best performance results on the test were at 0.528 which is less optimal.

Lilliana and colleagues [3] have classified Indonesian news with SVM techniques. With the concept of multiclass one-against-one (OAA), experiments performed with different SVM parameter values. 180 data from *kompas.com* news processed with SVM parameter Gamma from 1 to 1,5, and |C| value from 60 to 150. The minus of this research is they we're not clear about how to define the SVM parameter.

## III. METHODOLOGY

### (1) SVM For Classification

The support vector machine (SVM) has been reported as a discriminant classifier which is more accurate than most other classification models. The good generalization characteristic of the SVM is due to the implementation of Structural Risk Minimization (SRM) principle, which entails finding an optimal separating hyper-plane as illustrated in Eq. (1), thus guaranteeing the highly accurate classifier in most applications [11].

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \quad (1)$$

In Fig. 1, there are two groups of data points, represented by ''o'' and ''●'', which are linearly separated by some hyper-planes. In fact, there are an infinite number of hyper-planes (the dashed lines) could be generated, but there is only one hyper-plane (the solid line) which could optimally separate the data points from different categories. This optimal separating hyper-plane is located between the maximal margin, where in Fig. 1, margin is represented as d1 + d2.

The nearest data points to the optimal separating hyper-plane are called support vectors (SVs). There is a certain way to represent the SVs for a given set of training data points, and the maximal margin can be found by minimizing $\frac{1}{2}\|\mathbf{w}\|^2$, as shown in Eq. (2).

$$\min\{\tfrac{1}{2}\|\mathbf{w}\|^2\} \quad (2)$$

By minimizing Eq. (2), training data points are separated and optimal separating hyper-plane can be configured with the constraint as illustrated in Eq. (3)

$$y_i.(\mathbf{w}.\mathbf{x}_i + b) > 1, \forall i \quad (3)$$

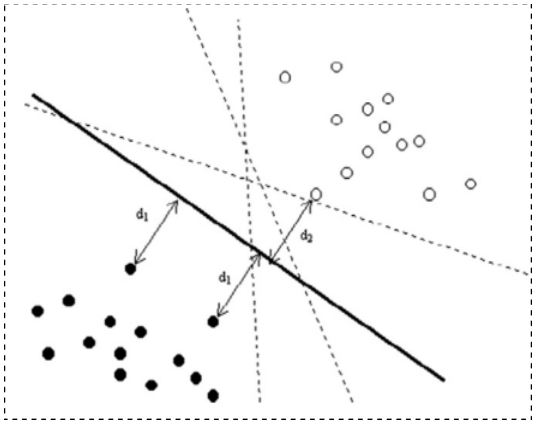


**Figure 1**. Optimal separating hyper-plane

### (2) Information Gain

The information gain measure is based on the entropy concept. It is commonly used as measure of feature relevance in filter strategies that evaluate features individually [12], and this method has the advantage of being fast. Let $D$ ($A_1, A_2, ..., A_n, C$), n ≥ 1, be a data set with $n$+1 attributes, where $C$ is the class attribute. Let $m$ be the number of distinct class values. The entropy of the class distribution in $D$, represented by *Entropy (D)*, is defined by Equation 5.

$$Entropy\ (D) = -\sum p_i * log_2(p_i) \quad (5)$$

Information gain (IG) is often used as a determinant of attributes in machine learning field. The IG of a term is measured by counting the number of bits of information extracted from a predicted category by the presence or absence of terms in a document. Mathematically can be written as follows [12]:

$$G(t)=-\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i) + P_r(t)\sum_{i=1}^{m} P_r(c_i|t)\log P_r(c_i|t) + P_r(t')\sum_{i=1}^{m} P_r(c_i|t')\log P_r(c_i|t') \quad (6)$$

Information gain (IG) measures the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. Concretely, it measures the expected reduction in entropy (uncertainty associated with a random feature).

### (3) SVM With IG

The method implemented in this paper is a hybrid method that combining SVM proposed algorithm and the Information Gain feature selection method. We started all process by download news data from some popular news portal in Bahasa like *kompas.com, detik.com, liputan6.com.*

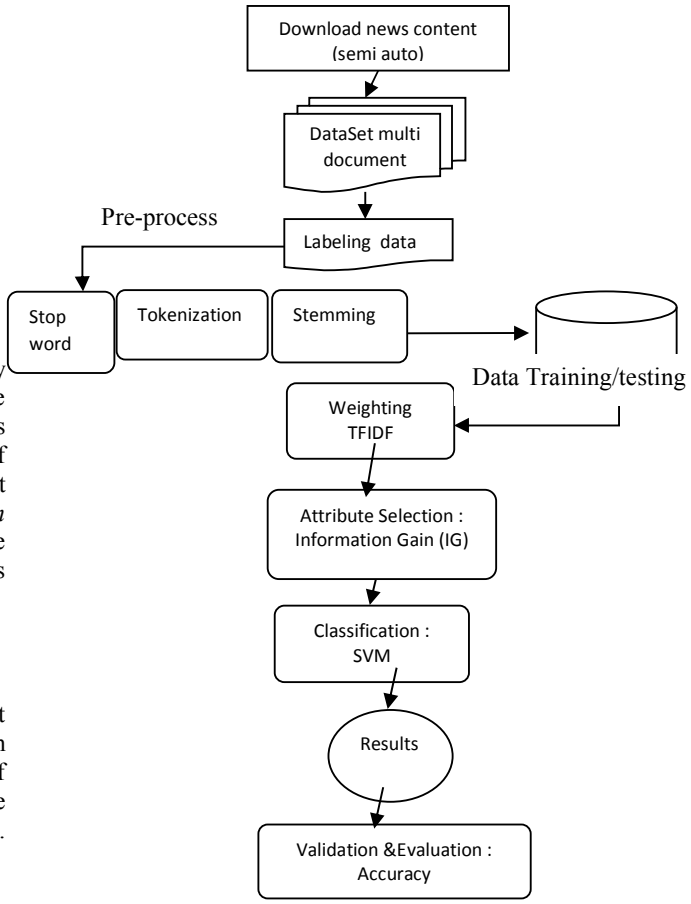


**Figure 2**. Flowchart of the classification

**Table 1**. Crawler web page URL pseudocode

| | |
|---|---|
| Crawling find active link | # Search data news algorithm<br>Reading inputurl<br># Desired category: [politics, business finance, national, entertainment, sport]<br>{ iteration start<br>Getting url root, url child<br>Finding text metadata/html, url, title.<br>If url is valid -> Collect url : title, content<br><html> tag.<br>Decode to saving .<br>Iteration stop }<br>Counting collection total<br>Performing parsing of files. |

We implemented crawling and scrapping technique for extract information from html documents. A logic pseudo code shown in Table 1 we used to break into line of codes. In this stage we do code using Python + Beautiful Soup framework. Files of data set then we give label by clustering into a folder. The name of the folder is in line with the pre-category from the domain of the online articles source.

Pre-processing stage is done with common steps in text classification that is stopword removal, stemming and tokenization. First of all is stopword removal by implemented Sastrawi open source library in Python framework [13]. We add some words in the code to remove English words that embedded between text. These terms is mostly html syntax characters that wasn't detected in web page scrapping step.

At this stage it will apply Enhanced Confix-Stripping Stemmer algorithm [14] to get the word base. This method has been commonly used as a word breaker with suffix, prefix, and affix with reference to the following model:

[AW+ [AW+ [AW+]]] Word Base [[ +AK][ +KK][ +P]]

With :
AW =Prefix KK = Possessive Pronoun
P = Particle AK = Suffix

The word in the next dataset will become a feature derived from breaking the sentence. This done by detecting punctuation marks like spaces, commas ",", dots "." and white spaces in each document. Since it does not take into account for semantic correlation between documents, the dot is considered as a token delimiter. Characters received as tokens are only letters with at least 3 characters, so non-alphabetical is ignored.

After the pre-process had done, next step was selecting feature or attribute using the Information Gain (IG) algorithm, as previously stated. IG was used to select the best attribute to be used in the classification process. Data learning process to build the model of Machine Learning was done using SVM algorithm. After the process of learning data is completed, classification phase start to build ML model. In SVM algorithm, we use linear kernel configuration. A complete process of our model is shown in Figure 2.

**(4) Evaluation**

Model testing using cross-validation technique, where this process divides data randomly into 10 sections. At the evaluation stage, the resulting value of the test is a classification of the pre-selected test data. To calculate performance on SVM implementation above, the measurement used is accuracy, and Precision. The scores is derived from the variables in a Confusion Matrix table.

## IV. EXPERIMENTS AND RESULTS

We use 5200 instances as dataset in our work. Every category distributed same amount of 1300 news from 5 sources. The detail of dataset shown in Table 2. After pre-processing step, feature is chosen by Information Gain selection. The Rank method of this IG, remove 73 features from 2035 to 1962 features. Our results from Weka output state that correctly Classified Instances is 5099 and uncorrectly Classified Instances is 101. The detail of this misclassified features figured out in Table 3. In result, our model give 98.057 % of successful categorization. Time taken to build model: 65.16 seconds. Compared without using feature selection, our SVM + IG improve 2.9% from 95.1 % to 98.06%.

**Table 2**. Data set Indonesia news

| kompas | detik | tempo | Liputan6 | tribunnews | newsportal/Category |
|---|---|---|---|---|---|
| 301 | 225 | 263 | 271 | 240 | Ekonomi |
| 465 | 200 | 154 | 419 | 62 | Entertainment |
| 523 | 166 | 188 | 219 | 204 | Olahraga |
| 329 | 220 | 478 | 221 | 52 | Teknologi |

**Table 3**. Confusion Matrix

| eko | entertainment | olahraga | tekno | <-- classified as |
|---|---|---|---|---|
| 1274 | 5 | 0 | 21 | Ekonomi |
| 9 | 1277 | 2 | 12 | entertainment |
| 2 | 4 | 1291 | 3 | Olahraga |
| 25 | 13 | 5 | 1257 | teknologi |

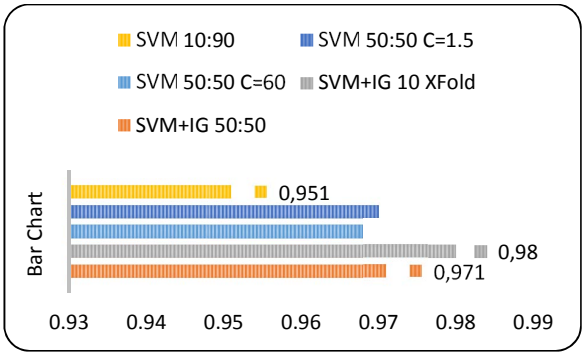


Figure 3. SVM News Classification Comparison

**Discussion**

Our experiments were done with some configuration with SVM as the basic algorithm. The first part of the classification uses a technique with no feature selection with a ratio of trainer data sharing and testing 50: 50. We also did experiment with 10:90 and 70:30 ratio. The second part with the same ratio but with the model we offer is SVM with Information Gain as the term selector. The results show how the evaluation of 10-cross validation is the best way with a result of 98,057 (98.06) % accuracy. Our experimental graph data can be seen in Figure 3 where some almost the same value score not

displayed on the Bar chart. SVM is an old technique in mathematics and computer study. Meanwhile, SVM still use in vast specific research like text classification, data mining classification, sentiment analysis and many others.

## V. CONCLUSION

From this work, text news classified by SVM using Information Gain shows improvement. The method trains the classifier by remove uninformative features. The contribution of this paper is the integration of the two methods which were never tested on Indonesian language. From experiments, the algorithm yielded up to 98.06 % accuracy in offline with an average of 65.16 seconds computational time. The experimental results have shown that this approach works well on a language with dissimilar structure like Indonesia language with satisfying accuracy. In the future, we would like to improve the algorithm by incorporating a parameter defining method algorithm in the training process to optimize SVM calculation process.

## REFERENCES


[1] J. Jaafar, Z. Indra, and N. Zamin, "A category classification algorithm for Indonesian and Malay news documents," J. Teknol. Sciences Eng., vol. 2, pp. 121–132, 2016.

[2] A. D. Asy'arie and A. W. Pribadi, "Automatic news articles classification in Indonesian language by using Naïve Bayes Classifier method," Proc. 11th Int. Conf. Inf. Integr. Web-based Appl. Serv. - iiWAS '09, p. 658, 2009.

[3] D. Y. Liliana, A. Hardianto, and M. Ridok, "Indonesian news classification using Support Vector Machine," Int. J. Comput. Electr. Autom. Control Inf. Eng., vol. 5, no. 9, pp. 1015–1018, 2011.

[4] D. Rahmawati and M. L. Khodra, "Automatic multilabel classification for Indonesian news articles," ICAICTA 2015 - 2015 Int. Conf. Adv. Informatics Concepts, Theory Appl., pp. 1–6, 2015.

[5] J. Xuand H. Jiang, "An Improved Information Gain Feature Selection Algorithm for SVM Text Classifier," Proc. - 2015 Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discov. CyberC 2015, pp. 273–276, 2015.

[6] Ş. Taşciand T. Güngör, "LDA-based keyword selection in text categorization," 2009 24th Int. Symp. Comput. Inf. Sci. Isc. 2009, pp. 230–235, 2009.

[7] V. Chandani, "Komparasi algoritma klasifikasi MachineLearning dan feature selection pada analisis sentimen review film," Universitas Dian Nuswantoro, 2014..

[8] P. Wira Buana, S. Jannet D.R.M., and I. K. G. Darma Putra, "Combination of K-Nearest Neighbor and K-Means based on Term Re-weighting for Classify Indonesian News," Int. J. Comput. Appl., vol. 50, no. 11, pp. 37–42, 2012.

[9] D. Ariadi and K. Fithriasari, "Klasifikasi Berita Indonesia Menggunakan Metode Naïve Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer," J. Sains Dan Seni ITS Vol. 4, No.2, vol. 4, no. 2, 2015.

[10] A. S. Wibawa and A. Purwarianti, "Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning," Procedia Comput. Sci., vol. 81, no. May, pp. 221–228, 2016.

[11] Y. Dong, Z. Xia, M. Tu, and G. Xing, "An optimization method for selecting parameters in Support Vector Machines," Proc. - 6th Int. Conf. Mach. Learn. Appl. ICMLA 2007, pp. 50–55, 2007.

[12] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," Proc. Fourteenth Int. Conf. Mach., pp. 412–420, 1997.

[13] https://github.com/har07/PySastrawi. Accessed on 12-07-2017 until 14-08-2017.

[14] A. Z. Arifin, I P. A. KertaMahendra, H. T. Ciptaningtyas, "Enhanced confix stripping stemmer and Ants algorithm for classifying news document in representation of textual," The 5th International Conference on Information & Communication Technology and Systems, pp. 149–158, 2007.