Text Classification Using Support Vector Machine for Webmining Based Spatio Temporal Analysis of the Spread of Tropical Diseases

Fatimah Wulandini[1], Anto Satriyo Nugroho[2]

Faculty of Information Technology, Swiss German University
German Centre Building 6F, BSD City – Indonesia 15321
Center for Information & Communication Technology, Center for the Assessment & Application of Technology (PTIK – BPPT)
Jalan MH Thamrin 8 BPPT 2nd bld. 4F, Jakarta – Indonesia 10340
[1]fatimah.wulandini@student.sgu.ac.id, [2]asnugroho@inn.bppt.go.id

***Abstract*— Tropical diseases such as Dengue Fever, Malaria and Bird Flu have become epidemic and particular problem in Indonesia. As the number of such cases increases, the availability of information regarding these diseases is important in order to help experts in taking proper actions. Meanwhile, web mining is one of significant technologies applied to extract information from the web. By using web mining, spatio-temporal information of tropical diseases will be collected from the internet. The objective of this study is to develop text classification system using Support Vector Machine to classify the Indonesian textual information on the Web. Proper classification for every downloaded text document helps the information extraction system to construct spatio-temporal analysis so then can be visualized. While Support Vector Machine has shown its capabilities for classifying text since it works well in high-dimensional data and avoids the curse of dimensionality problem.**

***Index Terms*—text classification, support vector machine, web mining, tropical diseases.**

I. Introduction

Tropical diseases such as Dengue Fever, Malaria and Bird Flu have become epidemic and particular problem in Indonesia. They spread rapidly from big cities to remote areas and anyone can be the victim. As the number of such cases increases, the availability of information regarding these diseases is important in order to help experts in taking proper actions and predict the pattern of the disease itself. Meanwhile, web mining has been broadly acknowledged as one of data mining techniques aims to gain information from data collected in the internet [3]. By using web mining, spatio-temporal information of tropical diseases will be collected from the internet.

This study has objective in developing text classification system which classified the Indonesian textual information on the internet. As part of datamining research project conducted in BPPTeknologi (Agency for the Assessment & Application of Technology) [8], this study plays important role in web mining system since it has significant contribution in making spatio temporal analysis. Proper categorization for every Indonesian textual document helps the information extraction system to construct spatio temporal analysis so then can be visualized.

Several studies have been conducted regarding text categorization. Experiment on web document categorization have been conducted by Goevert in [4]. The experiment applied based on probabilistic description-oriented representation of web documents and k-nearest neighbor classifier. Apte, Damerau and Weiss stated in [1] that machine generated decision rules able to compete with human performance in text categorization. In [7] Lewis and Ringuette evaluated the performance of text categorization using Bayesian classifier and decision tree learning algorithm, whereas Joachims proved that better result can be achieved by using Support Vector Machines [6]. Joachims also shows the statistical model of text classification in [5].

After introduction, section 2 will briefly explain about text categorization and Support Vector Machine. Methodology will be elaborated in section 3 while section 4 reports the result of the experiment.

II. Proposed Method

A.Text Categorization

Text categorization aims in classifying documents into predetermined fixed categories [6]. Transforming documents into an appropriate representation for the learning algorithm and the classification task is the first step in text categorization. Each distinct word w_i in documents which occurs for certain number of times is corresponded to a feature. The word considered as features if it appears in the training set at least 3 times and it is not *stop-word* (like "and", "or", etc). This model of representation leads to thousands of dimension features spaces which needs feature subset selection to improve

generalization accuracy and to avoid *overfitting*.

Text classification has 5 properties [5]. Firstly, it has high-dimensional feature space. If each word in the training documents considered as feature space, then there will be more than 50,000 attributes in a few thousand training example. The second property is that document has sparse vectors. If each document only contains a small quantity of distinct word, this means that document vector are very sparse. Third, text has heterogeneous use of terms and fourth it also has high level of redundancy. Between each document, there are still possibilities of its document vectors to overlap each other. In this case, the word in particular document are may be contained in other documents identified as another distinct category. Last property is frequency distribution of words and Zipf's law. According to Zipf's law, there is small number of words that occurs very frequently whereas most word occurs infrequently. Moreover, Zipf's law says that if one ranks words by their term frequency, the r-th most frequent word appears roughly 1/r times the term frequency of the most frequent words.

B.Support Vector Machine

Support Vector Machine (SVM) [9] has strategy to find the best hyperplane on input space called the structural minimization principle from statistical learning theory. Structural Risk Minimization means obtaining hypothesis $h(\vec{x})=\{\vec{w}\cdot\vec{x}+b\}$ described by a weight vector $\vec{w}$ and a threshold *b*, such that the lowest true error can be guaranteed, where true error of *h* is the probability that *h* will make an error on randomly selected examples.

Basically SVM is a linear classifier. It finds the hyperplane with maximum Euclidean distance to the closest training examples. The best hyperplane can be calculated with maximizing the margin $\delta=\frac{1}{\|\vec{w}\|}$.

$$\min_{\vec{w}} \tau(w)=\frac{1}{2}\|\vec{w}\|^2 \quad (1)$$

$$\forall_{i=1}^{n}: y_i(\vec{x}_i\cdot\vec{w}+b)\geqslant 1 \quad (2)$$

$$L(\vec{w},b,\alpha)=\frac{1}{2}\|\vec{w}\|^2-\sum_{i=1}^{l}\alpha_i(y_i((\vec{x}_i\cdot\vec{w}+b)-1)) \quad (3)$$

Obtaining maximum margin can be formulized in constraint (1) subject to (2) as Quadratic Programming (QP) problem. Constraint (6) is used to solve the problem through calculating its optimized value. The optimal value can be found when gradient L = 0, thus constraint (6) can be modified like the following constraint.

$$\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j \vec{x}_i\vec{x}_i \quad (4)$$

$$\alpha_i\geqslant 0(i=1,2,\ldots,l)\sum_{i=1}^{l}\alpha_i y_i=0 \quad (5)$$

Hence can be inferred that support vectors are α_i with positive value.

III. Model, Analysis and Design

A. Bahasa Indonesia

Bahasa Indonesia is used as national language in Indonesia. It has the root from Melayu language and has been modernized as its development through time. Bahasa Indonesia, or widely known as Bahasa, has its standard for writing and speaking which written in Common Guide of Indonesian Language spelling (*Pedoman Umum Ejaan Bahasa Indonesia yang Disempurnakan*) [11]. The guide explains usage of letter, punctuation, capital and italic letter, also writing of words as well as adaptive words.

In Bahasa, suffixes can be found almost in every word. Suffix is used to make derivation of word and can have different meaning depend on what and how the suffix is positioned. Suffix in Bahasa is divided into three types [11]; simple suffix, combined suffix and specific suffix. Adding suffix into word can alter the structure of the word itself [11]. For instance, word having the first letter *s*, if combined with prefix *me-* will modify the letter *s* with *ny* resulting *meny-*.

Moreover, adaptive word is foreign word that has been assimilated into Bahasa and widely acknowledged by Indonesian citizen. The word usually comes from Arabic, Sansekerta, Portuguese, Chinese Hokkien and Dutch. Adaptive word is used in daily conversation however misconception also often occurs. The most common mistake happen in writing since the standard one is not socialized well, thus make it inconsistent.

B. Tokenization

The first thing in processing documents is to fracture the stream of characters into words or tokens, often called tokenization [10]. Tokenization is complicated task for computer program since certain characters can be found as token delimiters. Delimiters are the character space, tab and newline while the characters ( ) < > ! ? ” are sometimes considered delimiters and may not be delimiters depend on the environment.

In this study, tokenization is generally done by breaking sentence into tokens and omitting the non-alphabet characters including numbers. All capital letters are converted into lower case so that tokens can be alphabetically ordered and treated equally.

C. Lemmatization

Lemmatization, often referred as stemming, is a task in converting collected tokens to a standard form. The purpose of lemmatization is to trim down the number of distinct type in a text corpus and increase the occurrences of some individual types [10].

Lemmatization, in this study, is done manually. The reasons are Bahasa has numerous rules and words in adaptive form may be written differently in each article. As for the last reason, computer program will treat the

particular words distinctively. Therefore, stemming is performed by giving index to words that have the same root.

The process of lemmatization is initiated with eliminating redundant words which occurs in each category (remaining-menyisakan) words to be occurred in one or at most two categories only. The words which are eliminated also include one's name, place and foreign words usage.

The next step is labeling words with number, or indexing. Words having same root are labeled with the same number. Indexing intents to reach a root form with no inflectional or derivational prefixes and suffixes. By indexing number of distinct words is reduced from over 11,000 to 3713 distinct words.

IV. Experiment

The experiment compares performance of SVM using polynomial kernel with 3 other conventional methods which are Naïve Bayes classifier, k-Nearest Neighbor classifier, and C4.5 Decision Tree.

The experiment is conducted on small scale of datasets with 3713 features and 360 instances. The instances are splited into 240 instances as training evaluation and 120 instances to test the method performance.

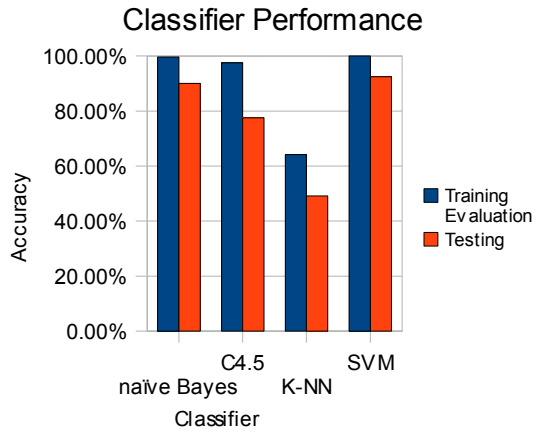

Figure 1. Classifier Performance

The above graph shows performance of each method in percent. SVM performs the best among other methods with 92.5% of accuracy. The SVM kernel used is polynomial with C = 250007 and E = 1. Meanwhile nearest neighbor classifier surprisingly performs the worst among other conventional methods with only 49.17% of accuracy using k = 11. The result opposes this [6] which stated that k-NN performed the best among other conventional methods. Naïve Bayes and C4.5 obtain 90% and 77.5% of accuracy respectively.

	Economy	Def & Sec	Education	Health	Sports	Politics
Economy	**17**	0	0	0	2	1
Def & Sec	1	**17**	0	0	1	1
Education	0	0	**19**	0	1	0
Health	0	0	0	**20**	0	0
Sports	0	0	0	0	**20**	0
Politics	0	1	1	0	0	**18**

Figure 2. SVM Confusion Matrix

Confusion matrix also shows that words in class label economy, defense & security and politics have similarities since misclassification mostly occurs on the respective class label. However, class label health and sports able to successively classify all test articles.

The outstanding performance of SVM shows that SVM still have better performance in classifying datasets with high dimensional features. This matches the characteristic of SVM which able to generalize well in high dimensional features and also omit the necessity for feature selection.

V. Summary

This paper develops Indonesia textual classification system for web mining based spatio temporal analysis of the spread of tropical diseases. The system is intended to classify downloaded Indonesia textual document from the Internet so then the information can be extracted.

The experiment result shows that SVM achieves good performance on Indonesian text classification similar to what it shows on English text classification. SVMs have the capability to generalize well in high dimensional feature spaces so that it requires no feature selection. Besides, SVMs are robust, outperforming other conventional methods in all experiments. Surprising result comes from naïve Bayes which shows good performance among other conventional methods. Language structure plays only minor role in this experiment since there is no difference on SVM performance when it is applied with Indonesia textual documents.

Finally, the experiment is still on process of optimizing. Future works will focus on increasing the datasets and exploring more on Indonesia language structure itself. The lookup table may be substituted with complete stemming algorithm in order to specify the words and distinguish each class label well making the prediction more accurate.

References

[1]Apte, C., Damerau, F., Weiss, S. M. Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems,* 1994, *vol 12*, p. 233-251.

[2]Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery,* 1998*, vol. 2*, p. 121-167.
[3]Chakrabarti, Soumen. Mining the web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers, 2003
[4]Gövert, N. A Probabilistic Description-Oriented Approach for Categorising Web Documents, 1999.
[5]Joachims, T. A Statistical Learning Model of Text Classification for Support Vector Machines. *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press,* 2001, p. 128-136.
[6]Joachims, T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features, 1997.
[7]Lewis, D. D. & Ringuette, M. A Comparison of Two Learning Algorithms for Text Categorization. *In Third Annual Symposium on Document Analysis and Information Retrieval,* 1994, p. 81-93.
[8]Prasetyo, B. et al. Desain Sistem Analisa Spatio-Temporal Penyebaran Penyakit Tropis Memakai Web Mining. *Inproceedings of Konferensi Nasional Sistem & Informatika,* 2008, p. 44-49.
[9]Vapnik, V. Statistical Learning Theory. Wiley, Chichester, GB, 1998.
[10]Weiss, M. Sholom et al. Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer, 2005.
[11]Pedoman Umum Ejaan Bahasa Indonesia yang Disempurnakan. *PT Gramedia Widiasarana Indonesia,* 1993.