Proc. of International Conference on Advance Computer Science & Information System (ICACSIS) 2010, pp.311-314, Bali, Indonesia, 2010

A Study on Text Classification for Webmining Based Spatio Temporal Analysis of the Spread of Tropical Diseases

Fatimah Wulandini, Iqbal Yasin, Anto Satriyo Nugroho, Bowo Prasetyo, Mohammad Teduh Uliniansyah, Vitria Pragesjvara, Made Gunawan, Gunarso, Ratih Irbandini, Dwi Handoko
Faculty of Information Technology, Swiss German University, Indonesia
Center for Information and Communication Technology
Center for the Assessment & Application Technology (PTIK – BPPT)
Jakarta, Indonesia
Email: wulan.fatimah@gmail.com, iqbal@sgu.ac.id, {asnugroho, praz, teduh, vitri, madegunawan, gunarso, irbandini, dwih}@inn.bppt.go.id

***Abstract*—The rapid growth of tropical diseases in Indonesia had led to countless number of victims. Experts had tried to overcome the problem by monitoring the spreading and collecting useful information regarding these diseases. Web mining is one technique to collect data information from the Internet. Spatio-temporal data of tropical diseases can be collected by using web mining so the useful information can be extracted for further analysis. The main objective of this study is to create a text classification system which classified the web document using several learning methods including naïve Bayes, nearest neighbor, decision tree and support vector machine (SVM) with Sequential Minimal Optimization algorithm. The classification is intended to construct a spatio temporal analysis for documents classified into health. The result showed that naïve Bayes and SVM-SMO achieve good performance (naïve Bayes: 95% and SVM-SMO: 92%). Multinomial distribution of naïve Bayes is able to normalize the length of document while SVM-SMO performed well in high-dimensional data.**

I. INTRODUCTION

The rapid growth of tropical diseases in Indonesia had led to countless number of victims and this become particular problem. Experts had tried to overcome this problem by monitoring the spreading and gathering useful information regarding these diseases. The internet can be utilized as main source for collecting the useful information. Textual Indonesian document from relevant source i.e. news could provide various information to help experts in taking proper action. Web mining is one of the techniques to gain information from data in the Internet [1]. By using web mining, spatio-temporal information of tropical diseases will be collected and extracted for further analysis.

The main objective of this study is to create text classification system, which classified the relevant Indonesian textual data gathered from the Internet. This study is part of web mining project conducted by Agency for the Assessment and Application of Technology (BPPT), Indonesia [2].

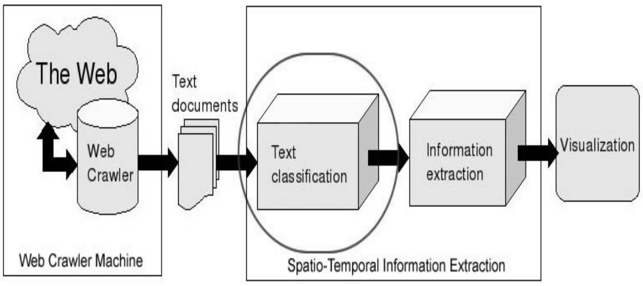


Figure 1. Illustration of the Web Mining Project.

The web-mining project that is conducted by BPPT team aims in monitoring the growth of tropical diseases in Indonesia. In the system, web mining is used to capture spatio-temporal information regarding the tropical disease from the Internet. The web mining system is developed in three parts. The first part is web crawler using open source software. The second part is spatio-temporal information extraction using text mining and the last part is visualization using Google Earth. The details of this web-mining project can be found in [2].

Study on text classification had been conducted by many researchers through years. Experiment based on probabilistic description-oriented representation and k-nearest neighbor had been conducted by Goevert [3]. Lewis and Ringuette evaluated the performance of text classification using Bayesian classifier and decision tree learning algorithm [4]. Apte, Damerau and Weiss stated in [5] that machine generated decision rule able to compete with human performance in text categorization, whereas Joachims stated in [6] better result is achieved using Support

Vector Machines. Also, in the previous study [7] Support Vector Machine was able to give good performance outperforming other classifiers.

In this study, the datasets is increased in total of 1050 articles while in previous study there were only 360 articles used as datasets. The new datasets will be converted into two types of vector, binary and term-frequency, aiming to evaluate the performance of both vectors. In addition, classifier with several parameters will also be tried on both datasets.

The next section will briefly recall the data preprocessing step as well as the detail of vector generation step. The result and analysis will be explained in section 3 while section 4 will summarize the study.

II. METHODS

A. Data Preprocessing

The data preprocessing step is initiated with tokenization process. Tokenization aims to fracture the stream of characters into tokens [8]. It is done by breaking the sentence into tokens while all non-alphabet characters are omitted. Furthermore, all capital letters are converted into lower case so that can be ordered alphabetically.

The next step is stemming the words with affixes to root words [8]. In this study, however, stemming is done manually since Bahasa Indonesia is relatively tricky. Stemming deals mostly with numerous rules of affixes, variation of writings and adaptive words. Stemming is initiated with eliminating redundant words which occur in every category leaving words to be occurred in one or at most two categories only. Then the words are labeled with number, which identifies each root words. Giving index intends to reach a root form with no derivational affixes. Indexing is able to reduce the size of dictionary from over 11,000 words to 3713 distinct words that next will be used for vector generation.

B. Vector Generation

After creating the lookup table, the next step is generating vector that represents the article. Vector generation aims to ease classifier for prediction. There are two types of vectors that are generated in this research, namely binary vector and term-frequency vector.

Binary vector is the simplest model. It stores the presence or absence of words in document corresponding to the lookup table. The entries will have ones or zeros, depending on whether the words are encountered in the document. The steps in creating binary vector are as followed:

1. Open the document and read on line at a time. Returns null if the end of file is reached, else returns the characters. Check whether the character is alphabet or not. Tokenized the sentence into words and keep tokenizing until end of file is reached.
2. Check the word to the lookup table whether it is encountered or not. If the word is present, attain its index number. Discard the words if they are not encountered in the lookup table.
3. Insert the index number into linked list. Check whether the index already exist in the list. Insert the index to the list if it is not exist and set the frequency to 1. Discard if the index already exists in the list. Do until all words are checked and sort the linked list.
4. Print the frequency stored in linked list. For index number whose words are not exist in the document, set the frequency to 0.

Binary vectors can work effectively when using large dictionary. However frequency can hold useful information for prediction [9]. Instead of zeros or ones, the actual frequency of occurrences of the words is used as the entries to the vector. The steps in generating term-frequency vector are almost the same with the binary one, except the third step. Instead of discarding the index, words frequency is incremented by 1. The rest steps are performed exactly the same with the binary vector construction.

III. RESULT AND ANALYSIS

In the experiment, both binary and term-frequency datasets were evaluated using naïve Bayes classifier [10], decision tree classifier [10], k-Nearest Neighbor classifier [10] and Support Vector Machine with Sequential Minimal Optimization algorithm [13]. The total instances for each datasets were 1050 articles that are divided into 700 articles as training set and 350 as testing set. The dataset was divided into seven categories namely Economics, Defense & Security (Def. & Sec.), Education, Health, Sports, Politics and Other. The performance of each classifier was measured using generalization error rate and precision-recall breakeven rate [12].

Figure 2 shows the classifier performance for binary vector. The lowest error rate is obtained by the modification form of naïve Bayes, called multinomial naïve Bayes. Meanwhile, the highest generalization error is achieved by nearest neighbor classifier. As expected, SVM-SMO able to compete with naïve Bayes with 7.71 % of error. Moreover decision tree as the representation of rule learner algorithm attains its performance moderately with 23.71 % of error.

High precision/recall-breakeven (PRBE) rate showed that the generated model was good while low PRBE rate indicates that the model was bad. The best model was generated by multinomial naïve Bayes while nearest neighbor classifier only produced moderately good model with only 73.1 %. SVM-SMO was also able to compete with multinomial naïve

Bayes since its PRBE rate was almost equal.

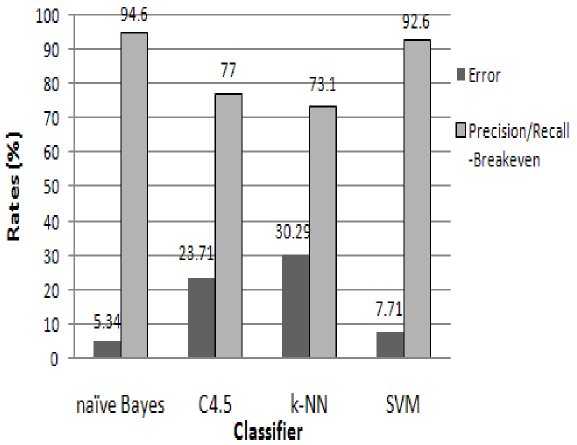


Figure 2. Classifier performance comparison for binary representation.

Since binary vector was simple, it worked effectively despite large size of dictionary. Words were treated as Boolean attributes making complicated computation could be avoided by classifier to find patterns for prediction. Therefore converting the article into ones and zeros vector gave advantages in classifying document.

Though binary vector gave advantages in its simplicity, words occurrences were potentially useful when determining the category of a document. However, not all learning algorithm could deal with continuous attributes and sparse vector. As observed in figure 3, nearest neighbor attained the highest generalization error with 58% while multinomial naïve Bayes succeed the task with only 4.57%. In contrast to binary vector, linear SVM-SMO performed better than non-linear SVM-SMO when dealing with frequency. Linear SVM-SMO was able to attain generalization error of 10.57%. Again, decision tree algorithm worked moderately with 20.57% of error.

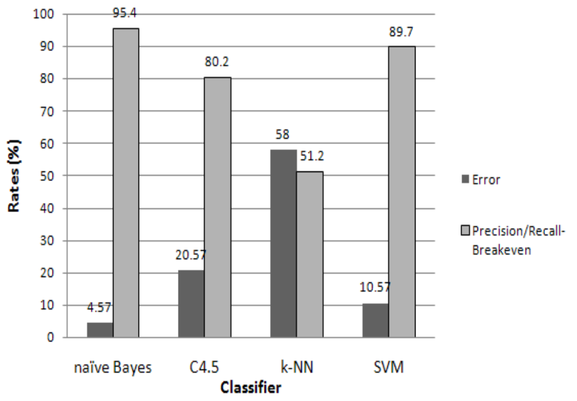
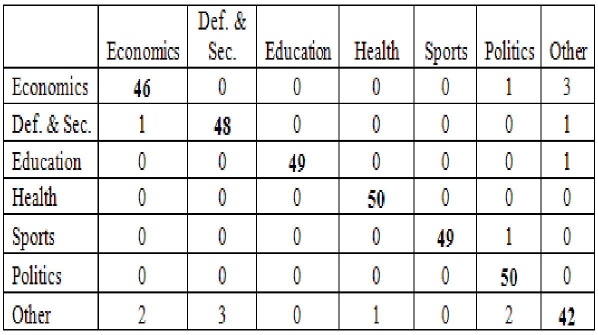


Figure 3. Classifier performance comparison for term-frequency representation.

Using term-frequency vector, naïve Bayes attained the highest PRBE rate meaning that it produced the best model. However, PRBE rate of SVM-SMO decreases to 89.7% showing that complex vector may lead to complex classifier computation. The improvement of performance was also attained by decision tree with increase to 80.2%. Nonetheless, nearest neighbor performed the worst, only 51.2%, with indication of overfitting in the generated model.

Both results show that SVM-SMO was underperformed by naïve bayes in binary and term-frequency representation. This is contrast to what Joachims showed in his paper [6] that SVM was able to outperform other learning algorithms in high dimensional vector space. The use of Bahasa Indonesia instead of English as dataset may give impact to performance of the learning algorithms used in this experiment. Unlike English, Bahasa had several unique characteristics that makes it relatively tricky to be categorized. In Bahasa there were usage of affixes, adaptive words and abbreviations that had not been standardized yet and hardly found in *Kamus Besar Bahasa Indonesia* (Big Dictionary of Bahasa Indonesia).

Moreover, multinomial naïve Bayes gave the best performance since it normalizes the length of document which often leads to better performance. The term-frequency vector achieved slightly better than binary vector since word frequencies could be accommodated using multinomial naïve Bayes. Although Support Vector Machine had shown excellent performance on text classification [6], SMO is still outperformed by multinomial naïve Bayes. The large size of dictionary resulted in a few number of words that lies on the wrong-side of hyperlane. SVM-SMO may be able to give better performance if the size of dictionary is reduced leaving only distinctive words from each category. Moreover, dictionary size also mades nearest neighbor classifier to produce high generalization error. There were several number of words which could be classified into two or more categories making it redundant.



	Economics	Def. & Sec.	Education	Health	Sports	Politics	Other
Economics	**46**	0	0	0	0	1	3
Def. & Sec.	1	**48**	0	0	0	0	1
Education	0	0	**49**	0	0	0	1
Health	0	0	0	**50**	0	0	0
Sports	0	0	0	0	**49**	1	0
Politics	0	0	0	0	0	**50**	0
Other	2	3	0	1	0	2	**42**

Figure 4. Term-frequency representation confusion matrix using multinomial naïve Bayes.

Word occurrences offered its advantages as well as disadvantages. Using word frequency as attributes might provide better calculation for prediction. Also it may yield to more compact solutions since the additional frequency was included to the same solution space as the binary data model. However it also produced more complex model, which led to longer computation time. Moreover the size of dictionary that was too large could also be restriction for some learning algorithm. Reducing the size of dictionary might improve the predictive performance and better training model.

The distribution for each class label can be seen in figure 4, which shows the confusion matrix. Class label Health and Politics were able to correctly classify all test articles. Class label Economic, Def. & Sec., Education and Sports share similar words since there were articles that are misclassified into each other. Furthermore, class label Other only able to classify 42 articles out of 50 articles showing that words from class label Other are vague to words from other class label.

IV. Conclusions

The text classification system developed in this study is for web mining based spatio-temporal analysis of the spread of tropical diseases. This system is intended to classify downloaded Indonesian textual document so then the information regarding tropical diseases can be extracted.

The result shows that binary dataset able to give better performance than term-frequency dataset. Simple representation of ones and zeros help the learning algorithm for prediction. Additional frequency was also useful for prediction yet it produces more complex model. However, for some learning algorithms, frequency of word occurrences provided slightly better result.

Among all learning algorithms, multinomial naïve Bayes achieved the best performance in both datasets. Normalization of the length of a document helped in classifying the document. Support Vector Machine with Sequential Minimal Optimization also showed very good performance since it had the capability to generalize well in high dimensional feature space. On the other hand, heavy overfitting was occurred when using nearest neighbor classifier leading to low generalization performance. The last algorithm used, decision tree, is only able to give moderate performance compare to other methods.

In the future the study may be focused on reducing the size of dictionary. Selecting keywords for every category and ranks the frequency of word occurrences also can be taken into account. Automation of word stemming with appropriate stemming algorithm for Bahasa Indonesia may substitute the usage of lookup table. Meanwhile implementing feature selection techniques may improve the data pre-processing performance, which then can be combined with the appropriate stemming algorithm for Bahasa Indonesia. Furthermore, the classification system will be integrated with Disease Spreading Monitoring System to complete the web-mining project.

References

[1] S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data, Morgan Kaufmann Publishers, 2003.

[2] B. Prasetyo et al. “Desain sistem analisa spatio-temporal penyebaran penyakit tropis memakai web mining”, Proceeding of Konferensi Nasional Sistem & Informatika, 2008, Bali, pp. 44-49.

[3] N. Goevert, M. Lalmas and N. Fuhr. “A probabilistic description-oriented approach for categorising web documents”, Proceeding of 8th International Conference on Information and Knowledge Management (CIKM), Kansas City, Missouri, 1994.

[4] D. D. Lewis and M. Ringuette, “A comparison of two learning algorithms for text categorization”, In Third Annual Symposium on Document Analysis and Information Retrieval, 1994, pp. 81-93.

[5] C. Apte, F. Damerau and S. M. Weiss, “Automated learning of decision rules for text categorization”, ACM Transaction on Information Systems, 1994, vol 12, pp. 233-251.

[6] T. Joachims, “Text categorization with support vector machines: learning with many relevant features”, Proceedings of the European Conference on Machine Learning, Berlin: Springer, 1998, pp.137-142.

[7] F. Wulandini and A. S. Nugroho, “Text classification using support vector machine for webmining based spatio temporal analysis of the spread of tropical diseases”, Proceedings of International Conference on Rural Information and Communication Technology, Bandung, pp. 189-192.

[8] S. M. Weiss et al. Text Mining: Predictive Methods for Analyzing Unstructured Information, Springer: New York, 2005.

[9] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann: San Francisco, 2005.

[10] P. –N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Pearson Education: Boston, 2006.

[11] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition”, Data Mining and Knowledge Discovery, Vol. 2, pp. 121-167.

[12] T. Joachims, Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms, Kluwer Academic Publishers, 2001.

[13] J. C. Platt, “Sequential minimal optimization: a fast algorithm for training support vector machines”, no. MSR-TR-98-14, April 1998.