

TEXT CLASSIFICATION TECHNIQUES USED TO FACILITATE CYBER TERRORISM INVESTIGATION

David Allister Simanjuntak, Heru Purnomo Ipung,
and Charles Lim
Faculty of Information Technology
Swiss German University
EduTown BSDCity, Tangerang, Indonesia 15339
Email: charles.lim@sgu.ac.id
Email: heru.ipung@sgu.ac.id
Email: david.simandjuntak@yahoo.com

Anto Satriyo Nugroho
Center for Information & Communication Technology,
Agency for Assessment & Application of Technology
BPPT 2nd building 4F, Jalan M.H.Thamrin 8
Jakarta, Indonesia 10340
Email: asnugroho@ieee.org

Abstract— rising of computer violence, such as Distributed Denial of Service (DDoS), web vandalism, and cyber bullying are becoming more serious issues when they are politically motivated and intentionally conducted to generate fear in society. These kinds of activity are categorized as cyber terrorism. As the number of such cases increase, the availability of information regarding these actions is required to facilitate experts in investigating cyber terrorism. This research aims to create text classification system which classifies the document using several algorithms including Naïve Bayes, Nearest Neighbor, Support Vector Machine (SVM), Decision Tree, and Multilayer Perceptron. The result shows that SVM outperforms by achieving 100% of accuracy. This result concludes the excellent performance of SVM in handling high dimensional of data.

Keywords—cyber terrorism, data mining, web mining, text classification, feature selection

I. INTRODUCTION

In this research, the author created the text classification system to detect which documents contain the information related to cyber terrorism. In addition, this research focuses on the text classification analysis that has the capacity to be utilized in Web mining. The analysis includes the performance comparison of Naïve Bayes, Nearest Neighbor, Support Vector Machine (SVM), Decision Tree, and Multilayer Neural Network Perceptron in the term of cyber terrorism.

After introduction, section 2 will briefly explain about cyber terrorism and web content mining. Methodology will be elaborated in section 3, while section 4 reports the experimental result, followed by section 5 which will conclude the result of the experiments.

II. CYBER TERRORISM AND WEB CONTENT MINING

According on the Dorothy.E.Denning's testimony before the Special Oversight Panel on Terrorism, Cyber Terrorism was defined as the following statement:

“Cyber terrorism is the convergence of terrorism and cyberspace. It is generally understood to mean unlawful attacks and threats of attack against computers, networks, and the information stored therein when done to intimidate or coerce a government or its people in furtherance of political or social objectives. Further, to qualify as cyber terrorism, an attack should result in violence against persons or property, or at least cause enough harm to generate fear...” [1]

There are three types of web mining knowledge: Web Content Mining, Web Structure Mining, and Web Usage mining [2.] Web Content Mining is the application of data mining techniques to unstructured or semi-unstructured text, typically HTML-documents. The utilization of Web content mining is the aim of conducting the experiments in this research paper. In order to mine the text, text classification is needed as the process of extracting interesting and useful patterns or knowledge from the web text, since text documents are unstructured data.

III. MODEL, ANALYSIS AND DESIGN

Data acquisition is the first phase of conducting this experiment. The data sets used in this research study is based on English textual document which is downloaded manually from the internet. In doing the experiment, *Holdout Method* was chosen by author in performing the text classification. Thus, the data sets were separated into two sections for the training set and test set, the distribution of the dataset is defined as follow:

- (i) Training Set, consists of 400 samples (200 Cyber Terrorism samples + 200 Non-Cyber Terrorism samples)
- (ii) Test Set, consists of 200 samples (100 Cyber Terrorism Samples + 100 Non-Cyber Terrorism samples)

After collecting documents as the data set there are actually three main phases within this research. These three phases are

text pre-processing, Training, and Classification. In text pre-processing phase author conducts:

- (i) **Tokenization**, it is a process of handling text document by breaking its stream of characters into words, or more precisely, tokens [4].
- (ii) **Feature Selection**, in feature selection phase, the author studied and created the list of terms (dictionary) related to cyber terrorism. In addition, the author also applied *Best First* algorithm in order to conduct the research and compare the performance of classifier.
- (iii) **Vector Generation**, there are two types of vector representation in this research, which are term-frequency vector and binary vector representation.

Hence, the result of the classifiers will be compared in order to find out which algorithm performs the best in relation to this research topic.

IV. EXPERIMENTAL RESULTS

The experiments were conducted to undergo in Weka 3.6 for Windows operating system. As the first step of conducting the research, the author did some parameter tuning in order to achieve its best result of classifier. This section explains the performance comparison of five classifiers based upon term-frequency (TF) and binary vector representation, as well as the feature selection (FS). The result of the performance is given follow:

1. Naïve Bayes

Naïve Bayes classifier works with the conditional independence assumption. It does not compute the class-conditional probability of each X, but only have to estimate the conditional probability of each X_i , given Y. The formula is defined as follow:

$$P(Y|X) = \frac{P(Y) \prod_{i=1}^d P(X_i|Y)}{P(X)} \quad (1)$$

Since $P(X)$ is fixed for every Y, it is sufficient to choose the class that maximizes the numerator term, $P(Y)$. [3] In experimenting using Naïve Bayes classifier there was not any parameter tuning, which was conducted by the author. Table I below shows the results of accuracy of Naïve Bayes classifier.

As it can be seen in Tab.I, Naïve Bayes performs better when using binary vector representation by achieving 97% of accuracy. In addition, by applying feature selection, the performance improves by 2%, from 97% to 99% of accuracy.

TABLE I. NAÏVE BAYES ACCURACY

TF	Binary	TF (FS)	Binary (FS)
87%	97%	97%	99%

2. K-Nearest Neighbor

Nearest Neighbor classifier is known as lazy learners, each example in this classifier represented as

a data point in a k dimensional space, where k is the number of attributes.

In conducting the experiment using the nearest neighbor classifier, the best performance is achieved by term-frequency vector representation while using $k = 5$, without attribute normalization, and no computation in distance weighting. It reaches its best performance by achieving 95.5% of accuracy. On the other hand, for binary vector representation nearest neighbor classifier performs best when the parameter was tuned using $k = 1$, without attribute normalization, and no computation in distance weighting by achieving 80% of accuracy.

In the experiment with feature selection, term-frequency vector representation attains its best performance when using $k = 9$, without attribute normalization, and no computation in distance weighting. Using this tuning, it achieves 94.5% of accuracy. Furthermore, for binary vector representation using feature selection, it performs best using $k = 1$, without attribute normalization, and no computation in distance weighting. By applying this tuning, this model achieves 95% of accuracy. According to the results in Tab.II, it can be concluded that, in this case, the k -Nearest Neighbor classifier performs better while using no computation in distance weighting and without normalization on the attributes.

TABLE II. K-NEAREST NEIGHBOR ACCURACY

TF	Binary	TF(FS)	Binary (FS)
95.5%	80%	94.5%	95%

3. Support Vector Machine

In process of learning, SVM introduces a new strategy to find the best hyperplane in the input space, through a strategy called Structural Risk Minimization [5].

In experimenting using Support Vector Machine classifier, the author sets the complexity parameter as 1 and using linear SVM, with the assumption that the input space can be linearly separable. By applying these parameter tunings, term-frequency vector achieves 99% of accuracy as the same as binary vector. On the other hand, using the same parameter settings, term-frequency vector representation outperforms by achieving 100% of accuracy. Furthermore, for binary vector representation both using with and without feature selection, has similar performance results. Table III below shows the performance of Support Vector Machine based on each vector representation and with or without feature selection.

TABLE III. SUPPORT VECTOR MACHINE ACCURACY

TF	Binary	TF (FS)	Binary (FS)
99%	99%	100%	98.5%

4. Decision Tree C4.5

Decision Tree is a method for generating a rule set from a decision tree, which is a simple yet widely used classification technique. In evaluating the result of Decision Tree C4.5 the parameter option that was tuned is to prune or unprune the leafs of the tree. This section presents the result of classification accuracy based on prune or unprune. As displayed below in Tab.4, the results in any condition, either term-frequency or binary vector representation (with or without feature selection) and pruned or unpruned, are the same.

TABLE IV. DECISION TREE C4.5 ACCURACY

	TF	Binary	TF (FS)	Binary (FS)
Pruned	95.5%	95.5%	96%	96%
Unpruned	95.5%	95.5%	96%	96%

5. Multilayer Perceptron Neural Network

Multilayer Perceptron Neural Network is a method of classifying data which is designed by an attempt to replicate biological neural systems [2]. In conducting the experiment with Multilayer Perceptron, the amount the weights are updated was set to 0.3, the momentum applied to the weights during updating was set to 0.2, and the number of periods to train through was set to 500. Classifying the text using Multilayer Perceptron only succeed while using feature selection approach due to the computational complexity issue in the classifier. Table V below shows the level of accuracy of Multilayer Perceptron. Term-frequency vector representation performs very well by attaining 99.5% of accuracy, while binary vector representation attains 95.5% of accuracy.

TABLE V. MULTILAYER PERCEPTRON ACCURACY

TF	Binary	TF (FS)	Binary (FS)
-	-	99.5%	95.5%

By evaluating the results of each classifier, apparently the use of feature selection in particular *BestFirst* algorithm leads into several advantages. It can be seen from the reduction of the significant number of attributes by applying feature selection.

In conducting the experiment based upon term-frequency vector representation there is reduction number of attributes from 819 to 37 attributes. It means that there is 95.5% of

dimensionality reduction. Besides, the experiment based upon binary vector representation feature selection performs by reducing 96.1% of dimensionality reduction. This percentage number represents the reduction from 819 attributes into 32 attributes. In order to give clearer presentation about selected attributes in experimenting using feature selection, Fig.1 below shows the list of terms (bag-of-words) being used in experimenting using feature selection

No.	Term	No.	Term	No.	Term
1.	against	14.	electronic	26.	sophisticated
2.	agencies	15.	experts	27.	system
3.	attack	16.	government	28.	team
4.	attacks	17.	governments	29.	terror
5.	bombing	18.	hackers	30.	terrorism
6.	bombs	19.	individuals	31.	terrorists
7.	communications	20.	intelligence	32.	threat
8.	computer	21.	network	33.	threats
9.	counter	22.	nuclear	34.	war
10.	cyber	23.	package	35.	warning
11.	cyberspace	24.	party	36.	weapon
12.	cyberterrorism	25.	security	37.	web
13.	digital				

Figure1. Bag-of-Words (Feature Selection)

V. CONCLUSION

It can be concluded this research has developed a proof-of-concept of a methodology to detect documents which contain information related to cyber terrorism using text classification techniques based on English textual document. In addition, by applying feature selection known as *Best First* algorithm it can avoid computational expensive and cut the execution time without decreasing the performance of the classifiers and even improve its level of accuracy.

Last, by comparing the result of each classifier, it shows that Support Vector Machine algorithm has the best by achieving 100% of accuracy based upon term-frequency representation with feature selection. This result proves that the capability of Support Vector Machine in high dimensional input space. As the future works in relation to this research is the used of TF-IDF (Term Frequency-Inverse Document Frequency) as the vector representation.

REFERENCES

- [1] Denning, Dorothy E. "CYBERTERRORISM. Testimony before the Special Oversight Panel on Terrorist Committee on Armed Services U.S. House of Representatives" *Georgetown University*, May 23, 2000. <http://www.cs.georgetown.edu/~denning/infosec/cyberterror.html>, accessed May 2, 2010.
- [2] J. Fürnkranz, Web Mining, in O.Maimon and L. Rokach (Eds.), *The Data Mining and Knowledge Discovery Handbook*, pp.899-920, Springer, 2005
- [3] Tan, P.-N., M. Steinbach, and V. Kumar. *Introduction to Data Mining*, Boston, MA: Pearson Education, Inc., 2006.
- [4] Weiss, S.M., N. Indurkha, T. Zang, and F. J. Damerau. *Text Mining Predictive Methods for Analyzing Unstructured Information*. New York: Springer, 2005.
- [5] Tsuda K. (2000). Overview of Support Vector Machine, *Journal of IEICE*, Vol.83, No.6, pp.460-466 (in Japanese)