UTM Computing Proceedings
Innovation in Computing Technology and Applications
Volume: 2| Year: 2017 |ISBN: 978-967-0194-95-0

Classification of Fraud Website Using Entropy Technique

Hazzaritta Binti Ya, Mohd Aizaini Bin Maarof

Department of Computer Science, Universiti Teknologi Malaysia, *Johor Bahru, Malaysia*

aizaini@utm.my

Abstract— Internet is a tool that provide information service where increasing of information leads to information issue. Web content was used by fraudulent to make crime activity by making their information look real and it able to deceive internet user. The impact from this attack, it makes victims loss their property. Web content consist the hyperlinks from the webpages and formatting the information in HTML tags has provide the feature which can be used in classification process. This project is proposed to use Entropy term weighting scheme as a feature selection to filtering the website and make the words use as a keyword that extracted from web document more significant. Support vector machine (SVM) will be used to make classification process and to define the accuracy performance by applying the entropy technique. Result shows the accuracy performance between category. The references cited will covered the theoretical part and guide the researcher to do their research more interesting.
Keywords—*feature selection, text classification, entropy technique, SVM supervise machine*

1. INTRODUCTION

The increasing of fraud web content in an internet was become problem of information filtering and it makes community worried to involve any offer from internet especially for investor to involve in any investment offer from organization that approach them through online. To make sure user are involve in legal investment, web filtering system have to be updated and developed.

The objective of designing the Web filtering is to control the content accessed especially when unlawful material is uploaded over the web [6]. The responsibility of web filtering is to specify content availability. Web filtering can be applied on several levels for example, it can be implemented in offices for employees, Internet Service Provider (ISP) to its clients, for students, and to his own computer. Web filtering has been applied in several countries at national level such as China, Cuba, Egypt, Iran, Burma, and Vietnam.

Figure 1 shows the most popular web filtering product that available in the market for the year 2012 (Internet, 2012). These web filtering products are limited in filtering because of they used the traditional technique [8] such as keyword matching, PICS and URL blocking.

Ranking	Product	Filtering Techniques		
		URL	Keyword	Content Analysis
1	Net Nanny	Yes	Yes	Yes
2	McAfee Safe Eyes	Yes	Yes	No
3	McAfee Safe Family	Yes	Yes	No
4	Pure Sight PC	Yes	Yes	No
5	Cyber Sitter	Yes	Yes	No

Figure. 1. World Top 5 Web Filtering System for Year 2010 (Internet, 2012)

In the above table, Net Nanny is the most effective web filtering software and it use content analysis so it was ranked as a higher software. It shows that the content analysis is included in web filtering.

These web filtering products that available in the market needs further improvement which make them less competencies, especially today's constantly changing web content [10]. The product that lack linguistic analysis will affects the accuracy.

Web filtering focus on the content of the document instead of subject related so web filtering will analyst the location and appearance of the web pages [3].

Web content filtering is one of web filtering technique has been used in this research. The advantage of these technique is they require short time for processing. Besides, web content is emphasizing toward dynamic web content and multimedia. Web content is always change and this technique will discuss the failure of their technique to define the suitable web pages [9].

Web content is a structure which consist of hyperlinks and the formatting information is a HTML tags format has provided the features which can be help for classification process.

Entropy Term Weighting scheme was used for feature selection of web content filtering. In text analysis, the language always redundant which have different words but have same meaning [11]. To define the importance word, statistical measure knows as term weighting method is used [2]. Through this technique, numerals value be established. Entropy is a probability analysis.

As a result, web filtering technology becomes the new trend which automatically classified webs based on the contents [10]. Figure 2 show the filtering general process.

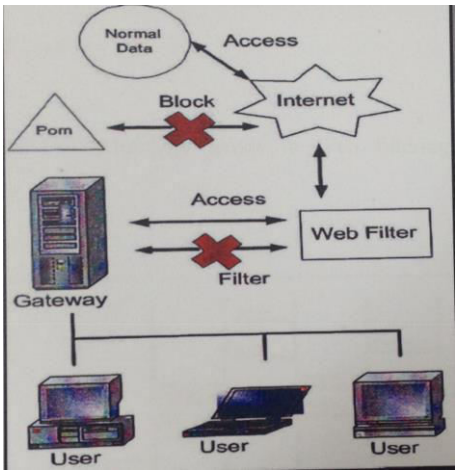


Figure 2. Web Filtering General Approaches

2. METHOD

Web filtering system automatically classified web sited based on their content using content analysis. Filtering web sites is an intelligent web content filtering that according to the semantic meaning of the image, pattern or text of the web content. Figure 3 shown the general view of web content analysis technique that contain several processes such as web data collection, pre-processing, data representation, feature extraction and selection, training, testing and classification [10]

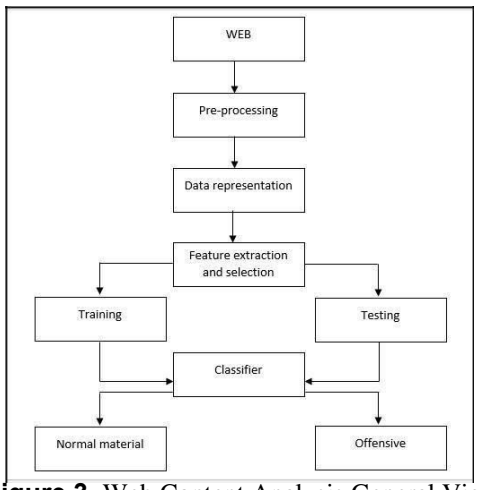


Figure 3. Web Content Analysis General View

2.1. Web Data Collection

Based on figure 3, the first process is data collection in the initial process to obtain the URLs that related with web pages for each category and from this data, the image and multimedia data will be excluded and only the text in HTML parsing will be taken out and used as a data.

Web crawler was used to get the related links based on the web page address provided and from all available web pages as shown in figure 4.

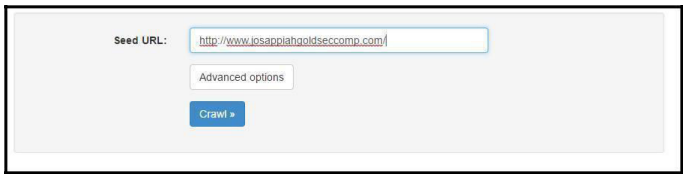


Figure 4. Web Crawler

Figure 5 shows the list of URLs that present in the web page brought by the crawler. URLs provided by crawler are many and each web page must extract the text from the particular web pages.

	URL	Page Title	Status	Meta Robots	Followed Ext. Links / Total Links	Rel Canonical	Potential SEO Issues
1	http://www.josappiahgoldseccomp.com/	JOSAPPIAH GOLD SECURITY COMPANY	200	N/A	0 / 23	Ok	
2	http://www.josappiahgoldseccomp.com/our-responsibility/	Our Responsibility – JOSAPPIAH GOLD SECURITY COMPANY	200	N/A	0 / 10	Ok	
3	http://www.josappiahgoldseccomp.com/about-us/	ABOUT US – JOSAPPIAH GOLD SECURITY COMPANY	200	N/A	0 / 10	Ok	
4	http://www.josappiahgoldseccomp.com/asset-security/	ASSET SECURITY – JOSAPPIAH GOLD SECURITY COMPANY	200	N/A	0 / 10	Ok	

Figure 5. List of URLs by Crawler

2.2. Pre-processing

Refer to figure 3, second step is to make pre-processing that consist of HTML parsing, stopping and stemming process. Figure 6 shows the HTML parsing process which using HTML parsing tool by fetching the URL and the tool automatically remove the HTML tags and only produce the text document.

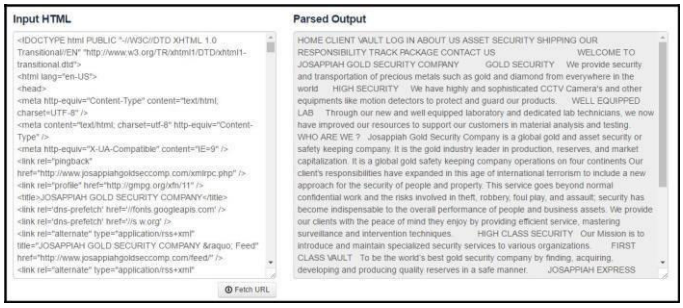


Figure 6. HTML Parsing Process

Figure 7 shows the stemming process to reduce the noise data which the words in document like their present root terms form.

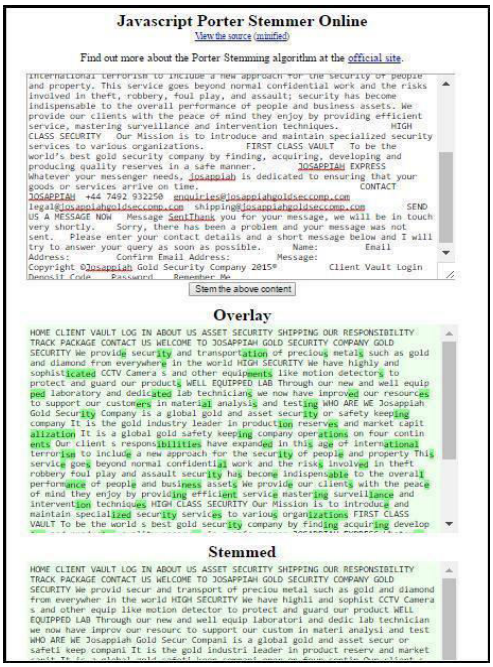


Figure 7. Stemmed Terms

Figure 8 shows the stopping process to remove the common words in web document such as “,”, “are”, “is” etc. and will saved as text document in the database. This step is important to remove the noise data and make feature selection process more ease.

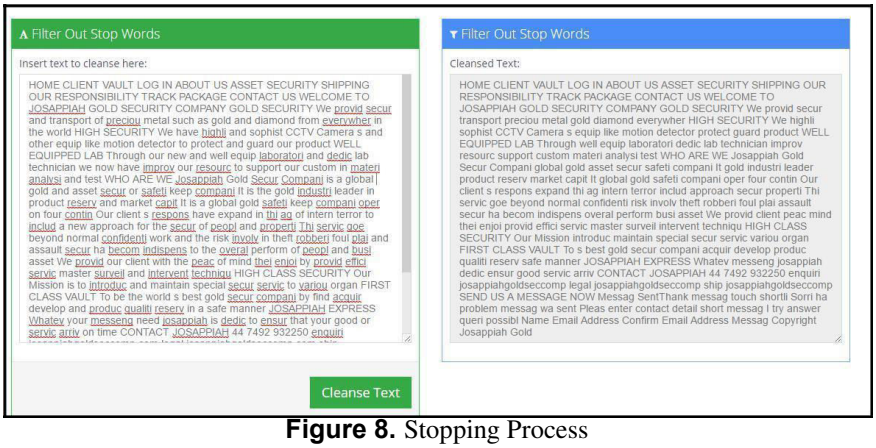


Figure 8. Stopping Process

2.3. Features Extraction and Selection

Feature selection technique build a powerful learning model to select a subset of relevant feature in machine learning. The main task of this technique is the features remove the possibility of selecting subset of input variable.

In terms of text categorize, it is ideal to represent the importance terms in text classification that called as term weighting scheme [10]. Term weighting scheme is a feature selection process in the text categorization which in certain condition of text categorization are less important and uninformative and by removing these condition does not affect the performance rating [4]. Make the classification of this condition also can increase accuracy and speeds up the process.

In this study, we use Entropy Term weighting scheme for feature selection. Entropy technique is based on probability analysis.

TF and DF is the elements that compute of term weighting schemes. TF is measure of how many times the term is occurrence in certain document. The formula of TF as follows.

$$x_i = \sum_{j=1}^{n} TF_{ij} \qquad (1)$$

DFi is the number of document contain the ith term in the set of collection. The formula of DF as follows.

$$x_i = \sum_{j=1}^{n} DF_{ij} \qquad (2)$$

The local and global term weighting that computes the weight are two important aspects of entropy.

$$G_i = \frac{1+\sum_{i=1}^{n} \frac{TF_{ij}}{F_i} \log\left(\frac{TF_{ij}}{F_i}+1\right)}{\log N} \quad (3)$$

Lij, jth, and ith are weight of document that represent to local time and for global time which is Gi represent the total weight of the ith term.

$$L_{ij} = \begin{cases} 1+\log TF_{ij} & (TF_{ij} > 0) \\ 0 & (TF_{ij} = 0) \end{cases} \quad (4)$$

$$x_{ij} = L_{ij} \times G_i \quad (5)$$

2.4. Training and Testing

SVM training is the process of learning that train a set of positive from negative examples that separating hyperplane [3]. Hyperplane is located in hyperspace that nearest point away from negative and positive examples and called as support vector.

The disadvantage is such as both testing and training size limit and speed performance. SVM classification process is very consuming and it's only applicable for binary classification.

3. CLASSIFICATION

Classification is to classified the input data into various groups. Web filtering is to separate forbidden from useful material and it needs the binary classification.

Two types of data to be classification are training and testing data and these two types will be applied in machine learning technique [9]. The most popular approach is Support Vector Machine (SVM).

The advantage of SVM text classifier which SVM can manage exponential numerous features and transformed space not requires to represent examples and two examples that contain similarity is required to effective.

4. RESULTS

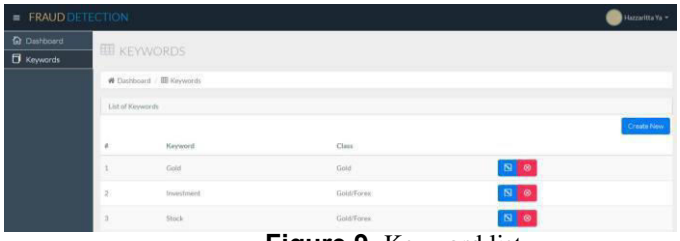


Figure 9. Keyword list

First, define the frequency for each keyword by using crawler that embedded in php code developed in Laravel framework as shown in figure 9.

Figure 10 shown list of keywords with their frequency for each document saved in database and be used as an input in Rstudio to define their frequency weight.

	Gold	Investment	Stock	Security	Shares	Forex	Financial	Foreign	Exchange Rates	Margin	Transaction	Currency
http://www.genesisminerals.com.au/australia.php	14	0	0	0	0	0	0	0	0	2	0	0
http://www.goldbridgecapital.com/	0	1	0	0	0	0	0	0	0	0	0	0
http://bobgoldpr.com/	6	0	0	0	0	0	0	0	0	1	0	0
http://www.goldmanreeves.co.uk/	0	0	0	0	0	0	0	0	0	0	0	0
https://www.optionrally.com	1	0	0	2	0	0	2	0	0	2	0	0
http://www.goldexasia.com/	2	1	0	0	0	2	1	0	0	0	0	0
http://www.icmtrading.com/	0	1	0	0	0	7	2	1	1	0	7	0
https://www.instaforex.com/ms/	5	1	1	0	0	89	4	0	0	0	28	0
https://www.netotrade.com/	0	0	0	0	0	0	0	0	0	0	0	0
https://www.sunbirdfx.com/	0	0	0	2	0	0	0	0	0	0	1	0
https://www.fxhq.com/	0	6	2	0	0	26	3	3	6	2	93	0
https://www.youtradefx.com/	0	0	0	2	0	0	0	0	0	0	1	0
http://www.gcitrading.com/	0	0	0	0	0	0	0	0	0	0	0	0

Figure 10. Data used in Rstudio

Figure 11 show the Entropy code development be used in Rstudio to define the weight for each frequency of keyword.

```
info <- function(CLASS.FREQ){
    freq.class <- CLASS.FREQ
    info <- 0
    for(i in 1:length(freq.class)){
      if(freq.class[[i]] != 0){ # zero check in class
        entropy <- -sum(freq.class[[i]] * log2(freq.class[[i]]))  #I calculate the entropy for each class i here
      }else{
        entropy <- 0
      }
      info <- info + entropy # sum up entropy from all classes
    }
    return(info)
  }
```

Figure 11. Entropy code development

Figure 12 shown the entropy code be implemented in Rstudio.

```
> Currency <- c("0", "0", "0", "0","0", "0","0", "1","0", "0","0", "16","0", "0", "0", "
0", "0","0", "0","0", "0", "1","7", "0","0", "2","0","0", "1","0", "0")
> freqs <- table(Currency)/length(Currency)
> info(freqs)
[1] 1.080424
> freqs
Currency
          0          1         16          2          7
0.80000000 0.10000000 0.03333333 0.03333333 0.03333333
>
```

Figure 12. Implement of entropy code development using Rstudio

After processing with R, each frequency for each keyword has their own weight as shown in figure 13.

Gold	Probability	Investment	Probability	Stock	Probability	Security	Probability
10	0.03333333	0	0.66666667	0	0.9666667	0	0.7
5	0.03333333	1	0.13333333	0	0.9666667	0	0.7
4	0.03333333	0	0.66666667	0	0.9666667	0	0.7
9	0.03333333	5	0.03333333	0	0.9666667	0	0.7
2	0.03333333	0	0.66666667	0	0.9666667	0	0.7
3	0.03333333	0	0.66666667	0	0.9666667	0	0.7
1	0.16666667	0	0.66666667	0	0.9666667	1	0.133333
1	0.16666667	0	0.66666667	0	0.9666667	0	0.7
1	0.16666667	0	0.66666667	0	0.9666667	2	0.1
1	0.16666667	1	0.13333333	0	0.9666667	0	0.7
1	0.16666667	0	0.66666667	0	0.9666667	0	0.7

Figure 13. Weight for each frequency

After finding the weight, these data used as an input for classification process. For classification process, the data must arrange according SVM classification format. As shown in figure 14, the data have been arranged nicely based on SVM classification format. In classification, the labelling must be declared for classified the document based on the labelling. This research use column **"**class**"** as a labeling

Class	Gold	Investmer	Stock	Security	Shares	Forex	Financial
Gold	0.033333	0.666667	0.966667	0.7	0.066667	0.533333	0.466667
Gold	0.166667	0.666667	0.966667	0.133333	0.9	0.533333	0.466667
Gold	0.033333	0.133333	0.966667	0.7	0.9	0.533333	0.466667
Gold	0.033333	0.666667	0.966667	0.7	0.9	0.533333	0.466667
Gold	0.033333	0.033333	0.966667	0.7	0.9	0.533333	0.133333
Gold	0.166667	0.666667	0.966667	0.7	0.9	0.533333	0.466667
Forex	0.633333	0.666667	0.966667	0.7	0.9	0.533333	0.2
Forex	0.633333	0.666667	0.966667	0.7	0.9	0.533333	0.466667
Forex	0.633333	0.666667	0.966667	0.7	0.9	0.533333	0.033333
Forex	0.633333	0.133333	0.966667	0.133333	0.9	0.033333	0.2

Figure 14. Data for classification

After transform the data, the data needs to be training and testing using Rstudio. To evaluate the data for testing and for training, SVM was used. Figure 15 shown the testing and training part

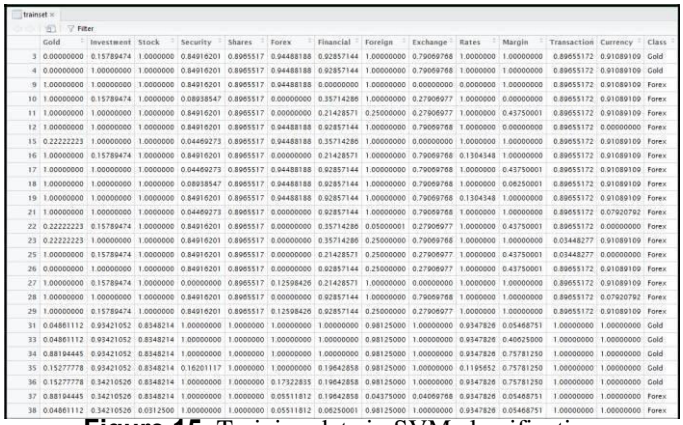
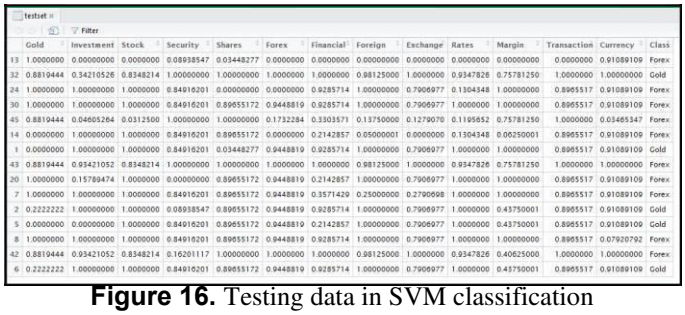


Figure 15. Training data in SVM classification



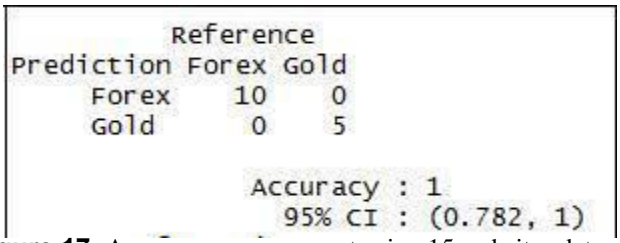
testset

Filter

	Gold	Investment	Stock	Security	Shares	Forex	Financial	Foreign	Exchange	Rates	Margin	Transaction	Currency	Class
13	1.0000000	0.00000000	0.0000000	0.08938547	0.03448277	0.0000000	0.0000000	0.00000000	0.0000000	0.0000000	0.00000000	0.0000000	0.91089109	Forex
32	0.8819444	0.34210526	0.8348214	1.0000000	1.0000000	1.0000000	1.0000000	0.98125000	1.0000000	0.9347826	0.75781250	1.0000000	1.0000000	Gold
24	1.0000000	1.0000000	1.0000000	0.84916201	0.00000000	0.0000000	0.9285714	1.0000000	0.7906977	0.1304348	1.0000000	0.8965517	0.91089109	Forex
30	1.0000000	1.0000000	1.0000000	0.84916201	0.89655172	0.9448819	0.9285714	1.0000000	0.7906977	1.0000000	1.0000000	0.8965517	0.91089109	Forex
45	0.8819444	0.04605264	0.0312500	1.0000000	1.0000000	0.1732284	0.3303571	0.13750000	0.1279070	0.1195652	0.75781250	1.0000000	0.03465347	Forex
14	0.0000000	1.0000000	1.0000000	0.84916201	0.89655172	0.0000000	0.2142857	0.05000001	0.0000000	0.1304348	0.06250001	0.8965517	0.91089109	Forex
1	0.0000000	1.0000000	1.0000000	0.84916201	0.03448277	0.9448819	0.9285714	1.0000000	0.7906977	1.0000000	1.0000000	0.8965517	0.91089109	Gold
43	0.8819444	0.93421052	0.8348214	1.0000000	1.0000000	1.0000000	1.0000000	0.98125000	1.0000000	0.9347826	0.75781250	1.0000000	1.0000000	Forex
20	1.0000000	0.15789474	1.0000000	0.0000000	0.89655172	0.9448819	0.2142857	1.0000000	0.7906977	1.0000000	1.0000000	0.8965517	0.91089109	Forex
7	1.0000000	1.0000000	1.0000000	0.84916201	0.89655172	0.9448819	0.3571429	0.25000000	0.2790698	1.0000000	1.0000000	0.8965517	0.91089109	Forex
2	0.2222222	1.0000000	1.0000000	0.08938547	0.89655172	0.9448819	0.9285714	1.0000000	0.7906977	1.0000000	0.43750001	0.8965517	0.91089109	Gold
5	0.0000000	0.00000000	1.0000000	0.84916201	0.89655172	0.9448819	0.2142857	1.0000000	0.7906977	1.0000000	0.43750001	0.8965517	0.91089109	Gold
8	1.0000000	1.0000000	1.0000000	0.84916201	0.89655172	0.9448819	0.9285714	1.0000000	0.7906977	1.0000000	1.0000000	0.8965517	0.07920792	Forex
42	0.8819444	0.93421052	0.8348214	0.16201117	1.0000000	1.0000000	1.0000000	0.98125000	1.0000000	0.9347826	0.40625000	1.0000000	1.0000000	Forex
6	0.2222222	1.0000000	1.0000000	0.84916201	0.89655172	0.9448819	0.9285714	1.0000000	0.7906977	1.0000000	0.43750001	0.8965517	0.91089109	Gold

Figure 16. Testing data in SVM classification

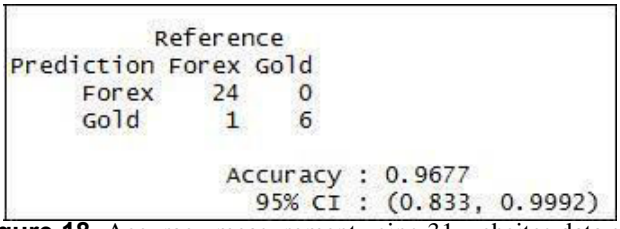
The experiment is conduct using different data set which 15 websites and 31 websites was tested. The information accuracy, used to evaluate the performance of entropy term weighting scheme as feature selection. The result of accuracy shown in figure 17 and figure 18.



```
          Reference
Prediction Forex Gold
     Forex    10    0
     Gold      0    5

               Accuracy : 1
                 95% CI : (0.782, 1)
```

Figure 17. Accuracy measurement using 15 websites data set



```
          Reference
Prediction Forex Gold
     Forex    24    0
     Gold      1    6

               Accuracy : 0.9677
                 95% CI : (0.833, 0.9992)
```

Figure 18. Accuracy measurement using 31 websites data set

5. CONCLUSION

From the research performed by Lee et all (2010), using entropy for web pages. When the number of words in collection set become bigger, the accuracy performance become less so entropy have been improving by control or improve the result from PCA to select the most relevant feature for the classification. After PCA is combine with feature vector in CPBF, the require manually selecting the most regular words in each class the terms should be weighted to find the measurement of entropy. The small number of document represent particular class in dataset may decrease the classification accuracy but with combination of PCA and CPBF, small number of web pages document, the accuracy of classification can be increase.

References

1. Boser, B.E., Guyon, I.M and Vaprik, V.N.(1992). A training algorithm for optimal margin classifiers. Paper presented at the Proceeding of the fifth annual workshop on Computational learning theory.
2. Chowdhury, G. (2010*). Introduction to modern information retrieval*. Facet publishing.

3. Du, A.N and Fang, B.X. (2004). *Comparison of machine learning algorithm in Chinese web filtering*. Paper presented at the machine learning and cybernetics, 2004. Proceeding of 2004 International conference on.
4. Efron, M., Zhang, J.and Marchionini, G. (2003). Comparingfeaturesselection criteria for term clustering application. Paper presented at the Proceeding of ACM SIGIR.
5. Eysenbach, G., diepgen, T.L., Gray, J.A.M., Bonati, M.,Impicciatore, P.,Pandolfini, C. and Arunachalam, S. (1998). Towards quality management of medical information on the internet: evaluation, labelling. And filtering of information Hallmarks for quality of information Quality on the internet assuring quality and relevance of internet information in the real world.Bmj, 317(7171), 1496-1502.
6. Gupta, S.,Kaiser, G.,Neistadt,D.and Grimm, P.(2002).DOM-based content extraction of HTML documents. Paper presented at the proceeding of the 12th international conference on World Wide Web.
7. Haselton, B.(2000). Study of average error rates for censorware programs. Retrieved octorber,25,2000.
8. Joachims, T.(1998).Text categorize with support vector machine. Learning with many relevant features. Machine learning : ECML-98, 137-142.
9. Lee,P.Y., Hui, S.C.andFong, A.C.M.(2002). Nueral network for web content filtering. Intelligent system. IEEE, 17(5), 48-57.
10. Lee, Z. S., Maarof, M.A., Selamat, A.and Shamsuddin. S.M. (2008). *Enhance Term Weighting Algorithm as feature selection technique for illicit web content classification*. Paper presented at the intelligent system design and application, 2008. ISDA'08. Eight international conference on.
11. Schulze, B. M. (2000). Automatic language identification using both N-gram and word information: Google paten