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

Data Analytics and Classification of Fraud Website

Nur Aina Shahira Binti Mazalan, Anazida binti Zainal

Faculty of Computing, Universiti Teknologi (UTM), Malaysia
aina.mazalan94@gmail.com, anazida@utm.my

Abstract. Almost everything is connected to the Internet nowadays. From a simple communication between two people to a large financial transaction involving two huge corporations, all is using the network as a medium to connect to each other. The existence of the Internet has made society depend on it more than the expected. People have been accessing Internet for many purposes such as to entertain themselves, to communicate with others, to search for information, and for commerce their products or services. All these stated purposes can be achieved by visiting a website – an online place that fulfils our needs. Unfortunately, this situation has leads to the increasing cases of fraudulent that relate to the website. People who fallen to the trick of the fraud website will face lot of consequences such as loss of personal information, financial loss, and many more. Organizations that operate using websites also face the same threat but with higher risk and bigger impact. Therefore, to avoid such thing from happening, a proper action needs to be taken. This study has proved to be able to classify web content into selected categories using keywords to classify fraud forex trading website, fraud gold investment website and unlicensed website using Support Vector Machine (SVM). Finally, the results obtained are presented graphically for easier understanding.

Keywords: Component, forex, gold, fraud, website, keywords, tf-idf, svm.

1 Introduction

Anything that uses the internet is exposed to cybercrime. All types of website possess their own risk of involving in cybercrime. The latest business trend is using online website. Online commerce or also known as e-commerce comes with both benefits and risk. Risk that e-commerce has are fraud, damaged customer and partner relationships, loss of intellectual property, unforeseen costs, public relations failure and business disruptions (Gartenberg, 2004). Categories of websites like information, communication and entertainment websites are also expose to different type of cybercrime such as identity theft, attacks on computer systems and illegal or prohibited online content. Online system such as websites is highly risky because the origin of the websites is unknown to the user. User cannot determine whether the website belongs to a licensed company or not. Novice user can easily be fooled by fake website. Fraud by using website is closely related to email because fraudulent websites is distributed by e-mail. This technique is called phishing. Phishing is a fraudulent e-mail that attempts to get you to divulge personal data that can then be used for illegitimate purposes. (Aryan et al., 2013). A lot of anti-phishing toolbars have been release by software vendors and companies to counter this threat (Yue et al., 2007). Researches Zhang et al. (2007) have developed CANTINA. CANTINA stands for A Content-Based Approach to Detecting Phishing Web Sites that focus on detecting phishing websites to using content based approach. TF-IDF algorithm is used to filter information.

2 Litrature Review

This section provides an overview of the literature review in the domain of Information Retrieval (IR). Cybercrime is a worldwide problem that is occurring every single day. Currently there are nearly 2 billion internet users and over 5 billion mobile phone connections worldwide. With that amount of user and counting cybercrime is on the rise. Cybercrime occurs in various ways depending what the criminals are after.

Fraud and financial crimes is one of the most common cybercrime occurring nowadays. Fraud is wrongful or criminal deception intended to result in financial or personal gain while online fraud or internet fraud is fraud with the use of Internet services, or software with Internet access. Financial crime is crime that is done to obtain money illegally. Fraud and financial crime are closely related to one another. Fraud in context of cybercrime is done using the internet as a medium to extort money out of victims either by obtaining victim's personal information or by tricking victim into paying money to fraudulent websites. Money is not the only motive for fraud but mostly the reason is money.

There are a lot of fraud detection existed that is credit card fraud detection, computer intrusion fraud detection and telecommunication fraud detection. The technique use for each of the detection is different. There is also detection for voting irregularities, criminal activities in e-commerce, insurance claims fraud, warranty fraud and abuse and health card fraud (Kou et al., 2004). All of the mention fraud detection specialized detection made for their domain. Among the technique use for the detection are outlier detection, neural network, data mining, model based reasoning and so on.

3 Rsearch Methodology

The research framework will outline all the necessary step to obtain the research objective. The framework is divided into three main phase based on the research objective. See Fig 1 of Research Framework.

Phase 1: Keyword Identification

Keyword Identification phase will focus on data collection and data cleaning. Input for this phase is list of websites and the output will be list of keywords. First of all, data from government and non-government agencies webpage will be collected. The webpage listed all the URLs that are suspicious or involve in fraud cases. Data cleaning is pre-processing data, this step is needed to obtain data that can be used in the next phase. Pre-processing are HTML parsing, Stopwords removal and Stemming process. Clean data as a result of all these processes will be evaluate to find keywords for forex and gold website.

Phase 2: Term Weighting Scheme and Classification Data

The outcome of phase 2 is classified websites of different category that is gold investment fraud, forex trading fraud and unlicensed business. Since the output of phase 1 is keywords, term weighting scheme is needed to calculate the weight of each term. Term weighting scheme use is Term Frequency Inverse Document Frequency (TFIDF). What Tf-idf do is decrease weight for commonly used words and increase weight for words that are not commonly used. After keywords go through Tf-idf, each of them will have a weight corresponds to the term. The data that consist of website list, keyword and weight of keyword will be divided into two sets, that is training data and testing data. The data is divided into one third (1/3) for testing and the rest are for training.

In the classification phase the expected outcome is classified websites. Support Vector Machine (SVM) is chosen to classify websites into categories. SVM will be trained with training data from phase 2. After the SVM learned from the trained data, testing data passed to SVM to classify websites based on the trained data given. After SVM is trained with train data a model of classifier is created. Testing data is passed to the model to categorized the testing data. The data will be categorized whether it is forex fraud, gold fraud or unlicensed. The output for phase 2 is classified website based on either one of the category.

Phase 3 : Visualization

The last phase will focus on visualization. Visualization will be made using proper tools and programming language. The outcome for this is a graphical dashboard that will showcase the overall research outcome. Visualization will be done using R programming language and Tableau. Data is analyzed using the R studio to generate appropriate graph or bar chart. Tableau software can create an interactive dashboard to better visualize the output. Visualization makes information easier to be understand by everyone.

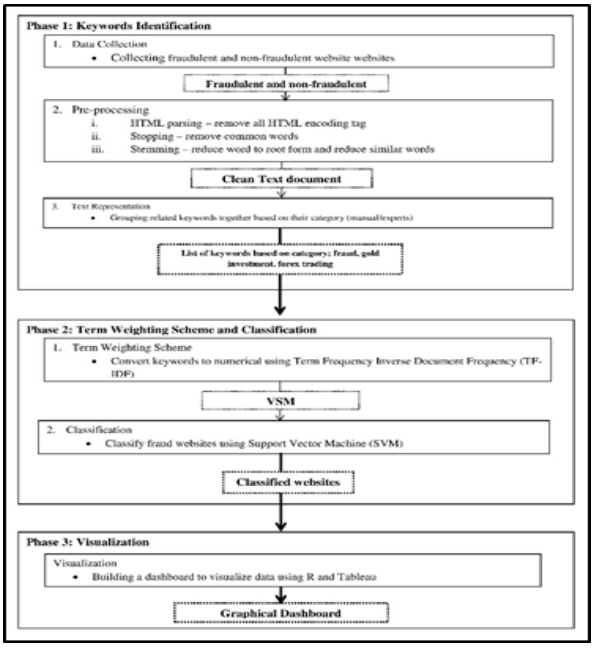


Figure1. Research Framework

4 Research Design and Implmentation

In this research, data collection begins with list of websites obtain from government agencies website. The list is put together in the website to warn people websites that are suspicious and highly related to fraud. Table 4.1 is the list of government and non-government agencies webpage where list of URLs is obtained.

The Financial Services Authority (FSA), UK - http://www.fsa.gov.uk/doing/regulated/law/alerts/unauthorised-firms Security Commission Malaysia (SCM)

i) https://www.sc.com.my/list-of-unauthorised-websites-investment-products-companies-as-referred-by-foreign-regulators/

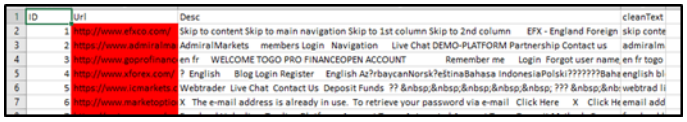
ii) https://www.sc.com.my/list-of-unauthorised-websites-investment-products-companies-individuals/

International Organization of Securities Commissions (IOSCO) - https://www.iosco.org/investor_protection/?subsection=investor_alerts_portal

Data are extracted using online tools that is Try jsoup (https://try.jsoup.org/). This website comes with HTML Parsing feature which discard all the HTML tag that exist in the extracted web content.

Pre-processing

Since online tool Try jsoup are used to extract websites content first step in pre-process has been done, that is HTML parsing. The next step to obtain an even cleaner dataset is to remove stop words, numbers, and stemming process. Figure 2 shows the arrangement of data that will be use.



	ID	Url	Desc	cleanText
2	1	http://www.efxco.com/	Skip to content Skip to main navigation Skip to 1st column Skip to 2nd column EFX - England Foreign	skip conte
3	2	https://www.admiralma	AdmiralMarkets members Login Navigation Live Chat DEMO-PLATFORM Partnership Contact us	admiralm
4	3	http://www.goprofinanc	en fr WELCOME TOGO PRO FINANCEOPEN ACCOUNT Remember me Login Forgot user name	en fr togo
5	4	http://www.xforex.com/	? English Blog Login Register English Az?rbaycanNorsk?eštinaBahasa IndonesiaPolski??????Baha	english bl
6	5	https://www.icmarkets.c	Webtrader Live Chat Contact Us Deposit Funds ?? ??? &nb	webtrad li
7	6	http://www.marketoptio	X The e-mail address is already in use. To retrieve your password via e-mail Click Here X Click He	email add

Figure2. Arrangement of Raw Data in Excel

Now that the raw data is arranged nicely, the next step of data cleaning is proceeded. The rest of the cleaning process is done in RStudio using tm package. Tm package has function tm_map that can do all the necessary cleaning of removing stop words, stemming, remove numbers and change the data to lowercase.

After cleaning process is done, in order to obtain keywords from all the document, frequency of all the words in the data are calculated. This process is done separately by category that is forex and gold to see what terms are significant to the category. From the word frequency, terms are selected by logic thinking

List of keywords shows the selected keywords use for this research. The selected keywords are based on the frequency of both categories. Since some of the keywords exist in both the categories, the next step will be executed by using the same keyword for both categories.

List of Keywords = ("trade","gold","forex","market", "account","invest","risk","island","price","open","compani","secur","world","spread","busi","manag","financi","start","support","asset","council","servic","demand","analysi","develop")

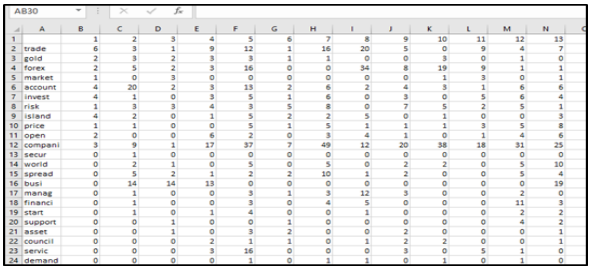
5 Result, Analysis and Diccussion

This chapter will explain details about selected term weighting scheme that is Term Frequency Inverse Document Frequency (TF-IDF) and classification technique using Support Vector Machine (SVM). Various ways had been tried to apply this term weighting scheme in order get the correct output. The result of these 2 phase will be discussed later on in the chapter.

I. Applying Term Weighting Scheme

Term Weighting Scheme is used as feature selection process. Term Weighting Scheme is used to calculate weight of a word in a collection of document. The selected algorithm for this process is Term Frequency Inverse Document Frequency (TF-IDF). Tf-idf is the cross product of two variable that is term frequency (tf) and inverse document frequency (idf). Calculation of tf-idf is done one by one referring to its formula.

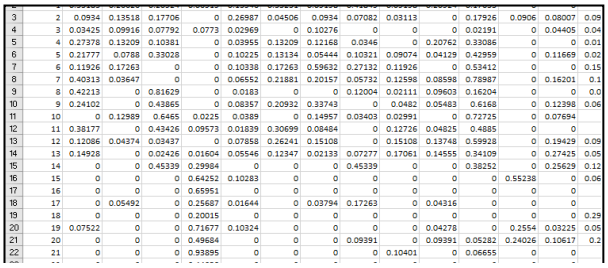
Calculation of tfidf is only apply to the selected keywords chosen in phase 1. The formula for this measure is tfidf(t,d,D)=tf(t,d)×idf(t,D) where t denotes the terms; d denotes each document; D denotes the collection of documents (Liu,2015). The outcome for this particular process is document term matrix that contain the term weight (Figure V.1).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		1	2	3	4	5	6	7	8	9	10	11	12	13
2	trade	6	3	1	9	12	1	16	20	5	0	9	4	7
3	gold	2	3	2	3	3	1	1	0	0	3	0	1	0
4	forex	2	5	2	3	16	0	0	34	8	19	9	1	1
5	market	1	0	3	0	0	0	0	0	0	1	3	0	1
6	account	4	20	2	3	13	2	6	2	4	3	1	6	6
7	invest	4	1	0	3	5	1	6	0	3	0	5	6	4
8	risk	1	3	3	4	3	5	8	0	7	5	2	5	1
9	island	4	2	0	1	5	2	2	5	0	1	0	0	3
10	price	1	1	0	0	5	1	5	1	1	1	3	5	8
11	open	2	0	0	6	2	0	3	4	1	0	1	4	6
12	compani	3	9	1	17	37	7	49	12	20	38	18	31	25
13	secur	0	1	0	0	0	0	0	0	0	0	0	0	0
14	world	0	2	1	0	5	0	5	0	2	2	0	5	10
15	spread	0	5	2	1	2	2	10	1	2	0	0	5	4
16	busi	0	14	14	13	0	0	0	0	0	0	0	0	19
17	manag	0	1	0	0	3	1	3	12	3	0	0	2	0
18	financi	0	1	0	0	3	0	4	5	0	0	0	11	3
19	start	0	1	0	1	4	0	0	1	0	0	0	2	2
20	support	0	0	1	0	0	1	0	0	0	0	0	4	2
21	asset	0	0	1	0	3	2	0	0	2	0	0	0	1
22	council	0	0	0	2	1	1	0	1	2	2	0	0	1
23	servic	0	0	0	3	16	0	0	0	3	0	5	1	0
24	demand	0	0	0	0	1	0	1	1	0	1	0	1	0

Figure3. List of frequencies of keywords in each document (tf)

From Figure 3 it can be seen that keyword “forex” and “gold” are among the most repeated keywords in the data set. It makes sense since it is the categories names. The keyword “company” and “trade” also appear a lot of times in both set of categories.



3	2	0.0934	0.13518	0.17706	0	0.26987	0.04506	0.0934	0.07082	0.03113	0	0.17926	0.0906	0.08007	0.09
4	3	0.03425	0.09916	0.07792	0.0773	0.02969	0	0.10276	0	0	0	0.02191	0	0.04405	0.04
5	4	0.27378	0.13209	0.10381	0	0.03955	0.13209	0.12168	0.0346	0	0.20762	0.33086	0	0	0.01
6	5	0.21777	0.0788	0.33028	0	0.10225	0.13134	0.05444	0.10321	0.09074	0.04129	0.42959	0	0.11669	0.02
7	6	0.11926	0.17263	0	0	0.10338	0.17263	0.59632	0.27132	0.11926	0	0.53412	0	0	0.15
8	7	0.40313	0.03647	0	0	0.06552	0.21881	0.20157	0.05732	0.12598	0.08598	0.78987	0	0.16201	0.1
9	8	0.42213	0	0.81629	0	0.0183	0	0	0.12004	0.02111	0.09603	0.16204	0	0	0.0
10	9	0.24102	0	0.43865	0	0.08357	0.20932	0.33743	0	0.0482	0.05483	0.6168	0	0.12398	0.06
11	10	0	0.12989	0.6465	0.0225	0.0389	0	0.14957	0.03403	0.02991	0	0.72725	0	0.07694	
12	11	0.38177	0	0.43426	0.09573	0.01839	0.30699	0.08484	0	0.12726	0.04825	0.4885	0	0	
13	12	0.12086	0.04374	0.03437	0	0.07858	0.26241	0.15108	0	0.15108	0.13748	0.59928	0	0.19429	0.09
14	13	0.14928	0	0.02426	0.01604	0.05546	0.12347	0.02133	0.07277	0.17061	0.14555	0.34109	0	0.27425	0.05
15	14	0	0	0.45339	0.29984	0	0	0	0.45339	0	0	0.38252	0	0.25629	0.12
16	15	0	0	0	0.64252	0.10283	0	0	0	0	0	0	0.55238	0	0.06
17	16	0	0	0	0.65951	0	0	0	0	0	0	0	0	0	
18	17	0	0.05492	0	0.25687	0.01644	0	0.03794	0.17263	0	0.04316	0	0	0	
19	18	0	0	0	0.20015	0	0	0	0	0	0	0	0	0	0.29
20	19	0.07522	0	0	0.71677	0.10324	0	0	0	0	0.04278	0	0.2554	0.03225	0.05
21	20	0	0	0	0.49684	0	0	0	0.09391	0	0.09391	0.05282	0.24026	0.10617	0.2
22	21	0	0	0	0.93895	0	0	0	0	0.10401	0	0.06655	0	0	

Figure 4. The TF-IDF weight of selected keywords

Next step is to train machine learning that is Support Vector Machine using the keywords weight data. Tf-idf count the weight based on how many times the word appeared in each document. Tf-idf are trying to avoid giving high scores to the term that appear a lot of time in just one document because that would be bias. Thus, the inverse document frequency plays an important role here. The main goal of tf-idf is to give high scores to the term that have high significance to that particular document and low scores to the term that occur less in the whole data set.

II. SVM Traning

To train SVM the data need to be arrange in a way that is easier for the coding to be run. SVM is use to classify data based on its category that is binomial classification. The data are modified a little to fit to the parameters of SVM. The table is modified by deleting the first column and add a new column Class at the right hand of table, Class is to represent

either forex, gold and unlicensed. The next step is training of SVM model. Dataset are separated to training and testing data. Before it is separated a randomized coding will randomly shift the row. Randomize is needed to obtain a distribute data, since the data are organized by category. The separation of training and testing data are done in one third (1/3) calculation. 1/3 of the data is used as testing data while the rest is use as training data. After the SVM model are trained the model will predict the category of forex, gold and unlicensed of the testing data.

```
svm.pred <- predict(svm.model, testset)
table(pred = svm.pred, true = testset[,26])
```

The prediction of svm are as follows

Table 1. Prediction of Testing Data

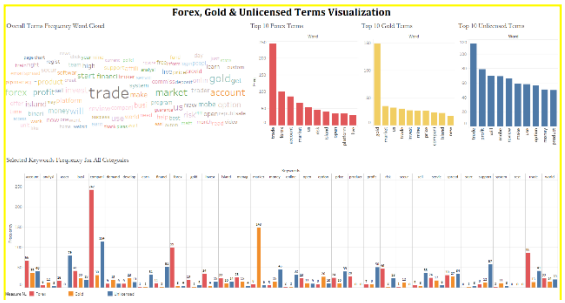
Pred\True	forex	gold	unlicensed	Total Prediction
forex	3	0	0	**3**
gold	0	3	2	**5**
unlicensed	0	0	3	**3**
Total True	**3**	**3**	**5**	

SVM model predicted that 3 of the documents are forex website, 5 are gold and 3 are unlicensed. 11 datasets are used for testing 24 for training. In this case, the SVM able to predict 83% precision for the testing data. To properly train SVM model a huge dataset of around hundreds need to be use.

III. Visualization

To achieve the third research objective that is to visualize the overall outcome of the research, a few graphical interpretations of output obtain is visualize. The first output is the keywords obtain to represent forex and gold categories is display in Figure 5.

Figure5. Word Cloud of Highest Frequency of Terms



6 Conclusion

This project aim is to develop fraudulent detection system using Support Vector Machine and further classify them into 2 categories forex trading and gold investment. The first objective is to pre-process the website content and to identify the fraudulent keywords. The website content has been processed manually to find the related keywords and from the frequency of some words appearances, the correct keywords have been identified. The second objective is to classify the website into different categories whether it is a gold investment fraud, forex trading fraud or unlicensed business. Support Vector Machine (SVM) is being used to categorize the website. Result that are obtained from the first objective which is the keywords are calculated to find the weight of each words relative to the total document and then processed using SVM. The machine learning algorithm is trained with separated data before being tested with another data. The result from the data testing show a very positive indication of a successful categorization of 83%. Finally, the result that has been obtained is visualize for better understanding. Using R programming, several visualizations have been created which has been discussed. In conclusion, all three objectives have been achieved and the output is successfully delivered.

References

[4] Marc Gartenberg, 2004. Retrieved on 25 March from http://www.computerworld.com/article/2574339/security0/e-commerce-and-web-presence--the-risks-and-threats.html.

[5] Mining Models (Analysis Services - Data Mining). 2016. Retrieved on 4 May from https://msdn.microsoft.com/en-us/library/cc645779.aspx

[6] Martin Porter The Porter Stemming Algorithm. 2016. Retrieved on 14 May from http://tartarus.org/martin/PorterStemmer/

[7] Sangwon Lee , Richard J. Koubek, 2010. The effects of usability and web design attributes on user preference for e-commerce web sites.

[8] Aryan Chandrapal Singh, Kiran P. Somase, Keshav G. Tambre, 2013. Phishing: A Computer Security Threat

[9] Yue Zhang, Jason Hong, Lorrie Cranor, 2007. CANTINA: A Content-Based Approach to Detecting Phishing Web Sites

[10] Lee, P. Y., Hui, S. C. and Fong, A.C. M., 2002. Neural Network for Web Content Filtering

[11] Approach to Detecting Phishing Web SitesChen, R.-C. and Hsieh, C.-H. (2006). Web page classification based on a support vector machine using a weighted vote schema. Expert Systems with Applications. 31(2), 427–435. Available at: http://www.sciencedirect.com/science/article/pii/S0957417405002307 [Accessed May 14, 2017