Content Based Fraudulent Website Detection Using Supervised Machine Learning Techniques

Mahdi Maktabar[1], Anazida Zainal[1](✉),
Mohd Aizaini Maarof[1], and Mohamad Nizam Kassim[2]

[1] Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia
momahdi3@live.utm.com, {anazida,aizaini}@utm.my
[2] Cyber Security Responsive Services, CyberSecurity Malaysia, Seri Kembangan, Malaysia
nizam@cybersecurity.my

Abstract. Fraudulent websites pose as legitimate sources of information, goods, product and services are propagating and resulted in loss of billions of dollars. Due to several undesirable impacts of Internet fraud and scam, several studies and approaches are focused to identify fraudulent Internet websites, yet none of them managed to offer an efficient solution to suppress these fraudulent activities. With this regard, this research proposes a fraudulent website detection model based on sentiment analysis of the textual contents of a given website, natural language processing and supervised machine learning techniques. The proposed model consists of four primary phases which are data acquisition phase, preprocessing phase, feature extraction phase and classification phase. Crawler is used to obtained data from Internet and data was cleaned to remove non-discriminative noises and reshape into desired format. Later, meaningful and discriminative patterns are extracted. Finally classification phase consists of supervised machine learning techniques to construct the fraudulent website detection model. This research employs 10-fold stratified cross validation technique in order to validate the performance of the proposed model. Experimental results show that the proposed fraudulent website detection model with cross validated accuracy of 97.67% and FPR of 3.49% achieved satisfactory results and served the aim of this research.

Keywords: Sentiment analysis · Text mining · Classification · Bag of words
Fraud detection

1 Introduction

In recent years, the expansion of the Internet, modern technology and ease of communication paved the way for fraudsters and criminals to conduct their fraudulent activities which results in the loss of billions of dollars worldwide each year [1]. Websites are great tool to access information, services and product, however like any other online service, they can be utilized by fraudsters to prey on victims and propagate fraud. Fraudulent websites are usually posing as legitimate online sources of information, goods, product and services [2]. Fraudulent investment websites such as foreign currency exchange (Forex), gold and other precious metal investment, Ponzi, pyramid schemes

A. Abraham et al. (Eds.): HIS 2017, AISC 734, pp. 294–304, 2018.
https://doi.org/10.1007/978-3-319-76351-4_30

and Multi-Level Marketing, online shopping and E-commerce website are the most common type of fraudulent websites [3, 4]. Due to various undesirable impacts of fraudulent websites, several studies and approaches are emerged to detect fraudulent websites. Despite the efforts, capabilities of these approaches are limited and they are unable to keep up with growth and diversity of fraudulent websites. Thus, the existing measures are fairly poor in terms of their ability to detect fraudulent websites [5]. The major challenges in fraudulent websites detection are: first, new generation of web technologies increased the complexity of web scrapping and limits the access of fraudulent website detection to the web contents. Second, diversity in types of web fraudulent activities (E-commerce fraud, MLM, Forex, etc.) challenge prescription of a global solution for fraudulent website detection. Third, viral growth of fraudulent websites, leads to obsolescence of static countermeasures and demands for a dynamic solution and lastly, fraudsters' efforts to disguise, mislead, block and bypass the fraudulent website detection models, leads to ineffectiveness of these models.

An effective fraudulent website detection model should be able to address these challenges and deliver accurate results within a reasonable time frame. This study proposes a fraudulent website detection model based on sentiment analysis of textual contents of a given website and supervised machine learning techniques. The proposed model was deployed in the Hadoop Big Data platform to ensure that it can keep up with vast amount and viral growth of fraudulent websites. It also employs content based machine learning techniques which increases its robustness against new types of fraudulent websites. Furthermore, a tailored crawling technique ensures the, availability of the required textual contents regardless of the web technology and type of the fraudulent website. The next section briefly describes the state-of-the-art fraudulent website detection models and their capabilities and limitation in more details.

2 Related Work

Content based fraudulent website detection methods rely on the website contents, components and metadata such as domain registration information, body text style features, HTML features, URL and anchor text features, image features, links, etc. in order to detect fraudulent websites. Various studies utilized these components to identify the legitimacy or fraudulency of a given website.

Le et al. [6] used lexical features of the URLs such as URL tokens, lexical and syntactic measures and domain registration information to identify fraudulent website. They claimed these features are immune against obfuscation techniques used by fraudsters. Different classification techniques including batch-learning, Support Vector Machine as well as online learning algorithms such as Online Perceptron, Confidence-weighted and Adaptive Regularization of Weights have been used in this study. Their evaluation results showed that Adaptive Regularization of Weights with accuracy of 96% outperformed other classification techniques.

In another study, Abbasi et al. [7] proposed a new fraudulent website detection systems based on statistical learning theory (SLT). Combination of textual, URL features, source code, images and linkage features were used to identify the fraudulent websites. Their evaluation results showed the accuracy of 96% using SLT in a dataset of 900 fraudulent websites. Another study by Abbasi et al. [7] attempted to detect fake escrow website using relatively similar features vector. Their proposed technique achieved accuracy of 98% using the kernel SVM classification technique.

Martinez and Araujo [8] proposed a scam website detection system using language model analysis. They applied a language model approach to different sources of information extracted from a given website to extract features for scam website detection. Their system relies on the hypothesis that two pages linked by a hyperlink should be topically related, even though this were a weak contextual relation. Three sources of features from the source page including Anchor Text, Surrounding Anchor Text and URL terms as well as three sources of information from the target page including Title, Content Page and Meta Tag were used to construct the feature vector. Their proposed model achieved F-score of 81% in detection of scam websites.

A study performed by Urvoy et al. [9] proposed an spam website detection model based on a style similarity measures of textual features in html source code. Jaccard similarity index has been used to measure the similarities. They also proposed a method to cluster a large collection of documents according to this measure. Their proposed technique is particularly useful to detect pages across different sites which sharing the same design. In a similar fashion, several other studies including [11–14] attempted to identify spam website using supervised machine learning techniques.

3 Methodology

Among the various components of a website which have been mentioned earlier, textual contents are the primary and presumably the largest element of a webpage. Textual contents are often explicitly imply information about the semantic of that webpage which can be employed as primary source of discrimination features for fraudulent website detection. Various types of discriminative features such as Lexical, Syntactic, Structural, Content-specific and Idiosyncratic can be extracted form textual information of a webpage. The proposed fraudulent website detection model employs Natural Language Processing (NLP) techniques which let us to go beyond rudimentary statistical features and extract the semantic features of the text through the use of natural language features [15, 16]. This study incorporate NLP techniques, textual features and machine learning techniques in order to construct fraudulent website detention model. The proposed model consists of four primary phases including data acquisition phase, preprocessing phase, feature extraction phase and classification phase. The following sections describe each phase in details. Figure 1 shows the proposed fraudulent website detection model framework.

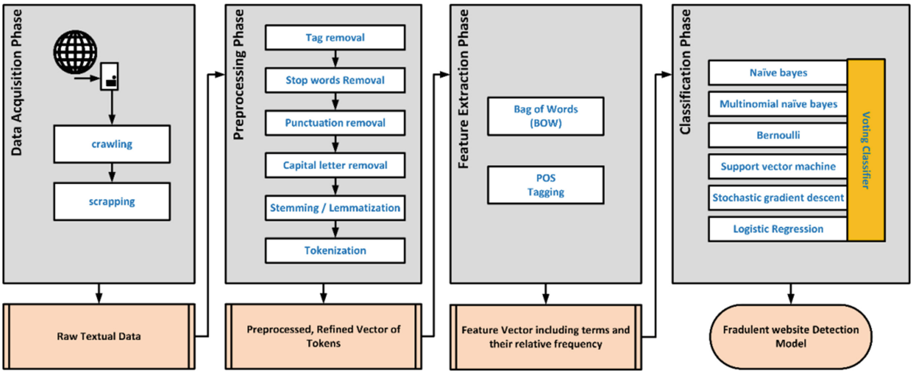

Fig. 1. The proposed content based fraudulent website detection model.

3.1 Data Acquisition Phase

This phase is aimed to extracts textural raw information and metadata from a given website. Web *crawling* and *scrapping* are two major operation which used in this phase.

Despite the seemingly simple operation of web crawling and scrapping, majority of the webpages devised restrictive measures such as browser ID check, CAPTCHA, IP monitoring, query limit, etc. to block the unknown crawlers and reduce unwanted load to their servers. Majority of the modern crawlers and scrappers are employing sophisticated techniques to bypass the restrictions imposed by websites administration. Also, diversity in web technologies, programming languages and styles make the web crawling more difficult than ever before. To tackle these issues, this study used Pyspider which provides highly customizable application, framework and libraries to *crawl* the target websites. Pyspider is an application framework for crawling web sites and extracting structured and unstructured data which can be used for data mining, information processing or historical archival. Web scraping software will automatically load and extract data from multiple pages of websites based on the given criteria and parameters. It is either custom built for a specific website or is one which can be configured to work with any website. In order to extract the textual data from a given webpage, scrappers are usually parse for the HTML tags and class objects such as <title> …</ title> , <p>… </p> , <div> …</div> which are usually indicate the present the textual contents. Our scrapper only collects textual data from the home page as well as first layer links. We believe this setup is adequate to collect enough textual data to identify website sentiment. Once textual contents of a given webpage scrapped, it will be stored in a dataset for subsequent mining and analysis operations. In this study total number of 430 website has been crawled and scrapped among which 257 denote the fraudulent website and the reset represents the legitimate ones. Textual content of these websites is then dumped in text files to facilitate the data cleaning and wrangling process.

3.2 Preprocessing Phase

Regardless of web scrapper precision, there are always significant amount of unwanted noise such as HTML or style tags incorporated into the actual textual data. These noise patterns which negatively affect the fraudulent website performance, should be removed prior to feature extraction and classification phase. This study employs several preprocessing techniques in order to remove unwanted noise and refine the textural data. The following explain each preprocessing technique in more details:

i. *Tag Removal*: Despite majority of the source code tags has removed from the textual data in scrapping phase, in some cases these tags can be sneaked into the textual data. These tags has been removed prior to any feature extraction process.
ii. *Stop Words Removal*: Stop-words are repeatedly occurring, insignificant words in a language but do not reflect any semantic of the documents. Articles, prepositions and conjunctions and some pronouns are common stop-words in English. Stop-words have been removed prior to any feature extraction operations.
iii. *Punctuation Removal*: Similar to stop words, punctuation marks are frequently occurring in the language but do not reflect any semantic of the given document. Punctuations have been removed prior to any feature extraction operations.
iv. *Capital Letters Removal*: Since capital letters do not affect the semantic of a given word, all capital letters throughout the document transformed to the small letters. This increases the uniformity of the document and cuts the redundancy.
v. *Stemming*: Stemming is the process of shrinking words into their stems or roots. For example "computer", "computing", and "compute" are reduced to "comput" and "walks", "walking" and "walker" are reduced to "walk". Martin Porter's stemming algorithm used in this study. Stemming increases the accuracy and reduces the size of the text block by factorizing the words to their stems.
vi. Tokenization: Tokenization can be performed at different scales such as words, phrases and sentence. Since this research uses BOW technique in feature extraction phase, we have employed tokenization operation in "word" scale.

After preprocessing operation the input raw textual documents are transformed to a refined vector of words which used in subsequent feature extraction phase.

3.3 Feature Extraction Phase

Feature extraction is the most important stage in fraudulent website detection using natural language processing. Various types of discriminative features which can be extracted form webpage body text. This research employed *Bag-of-Words* technique and *Part-of-Speech* tags in order to construct the feature vector and segregate the fraudulent websites. *Bag-of-Words* model is a simple yet efficient technique used in NLP. It makes a unigram model of the text by keeping track of the number of occurrences of each word in a given document. This technique creates a corpus with word counts for each data instance (document). Based on the average size of each document, word counts can be either absolute, binary (contains or does not contain) or sublinear (logarithm of the term frequency). For the purpose of this study, absolute word count generate superior results in comparison to other word count techniques. Figure 2 shows Bag-of-Words pseudocode.

```
Initialize feature vector bg feature= [0, 0, .....0]
for token in text.tokenize() do
    if    token in dict then
        token idx=getindex(dict, token)
        bg feature[token idx]++
    else
        continue
    end if
end for
return bg feature
```

Fig. 2. Bag-of-Words pseudocode.

Part-of-Speech tagging is the process of marking up a word in a given body of text corresponding to a particular part of speech, based on both its definition and its context for example nouns, verbs, adjectives, adverbs, etc. Several pre-trained part of speech taggers are publicly available online. One major reason to use Part-of-Speech tagging is to extract name entities and employ them in feature extraction process. In comparison to other parts of the speech, name entities provide significantly higher discriminative power which can be beneficial in feature extraction process. we have employed Brill rule based POS tagging technique in this study [17].

3.4 Classification Phase

This study employs ensemble of classifiers and a voting scheme to increase the reliability of the prediction. Combination of six classifiers including Naïve Bayes, Multi-nomial Naïve Bayes, Bernoulli, Logistic Regression, Stochastic gradient descent classifier and Support vector machine used to construct the ensemble classifier and voting scheme. Naïve Bayes is a probabilistic classifier based on Bayes' Theorem with an assumption of conditional independence of predictors. In training stage, it estimates the parameters of a probability distribution while in the testing stage it computes the posterior probability of that test sample belonging to each class. Despite simplicity, Naïve Bayes perform reasonably well for text classification purposes. Multinomial Naïve Bayes is a variant of Naïve Bayes which is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. Bernoulli is basically a Naïve Bayes classifier for multivariate Bernoulli models. Like Multinomial Naïve Bayes, this classifier is suitable for discrete data. The difference is that while Multinomial Naïve Bayes works with occurrence counts, Bernoulli is designed for binary/boolean features. Support Vector Machine is a two-class classifier which mainly intended to find the hyper plane which separates two classes with maximum marginal distance between them. It generates satisfactory result in small to medium scale dataset. Stochastic Gradient Descent (SGD) estimator implements regularized linear models with stochastic gradient descent (SGD) learning. The gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule (learning rate). Logistic Regression used to predict a binary outcome by using binomial Logistic Regression, or it can be used to predict a multiclass outcome by using multinomial logistic regression. The voting scheme is designed based on these classifiers. For example if four out of six classifiers vote for fraudulency of a website, the voting scheme classifies that website as fraudulent. 10-fold stratified cross validation technique is employed to validate the performance of the proposed model.

4 Results and Analysis

A total number of 430 website has been crawled and scrapped where 257 are fraudulent website and the rest are legitimate websites. 10 fold stratified cross validation technique was used to validate the results. Among the six classifiers used in this study, Logistic Regression and Multinomial Naïve Bayes Classifiers have outperformed others in terms of accuracy, F-score and FPR. Logistic regression with respective Accuracy, F-score, and FPR of 98.83%, 98.32%, 2.55% on 10 fold cross validation, generates the best results. Multinomial Naïve Bayes Classifier with respective Accuracy, F-score, and FPR of 99.41%, 99.11%, 1.64% marginally underperform the Logistic Regression and is among the best classifiers in this research. SVM classifier with respective accuracy, F-score and FPR of 88.37%, 87.38%, 14.01% is the worst performer in this study. Despite, Logistic Regression and Multinomial Naïve Bayes classifiers produce superior results than other classifiers in this study. Their performance are very much depends on the quality and type of input data and might fluctuate significantly as changes happen to input data. Voted classifier can significantly mitigate this issue and improve the robustness of the proposed model and increase the consistency of the results. The proposed voted classifier with respective Accuracy, F-score, and FPR of 97.67%, 97.25%, 3.49% though does not yields cutting edge results, but ensures consistent performance regardless of the fluctuations in input data. Table 1 also shows nearly perfect result in training stage which indicates high generalizability and discrimination power of the proposed feature vector. Table 1 shows the evaluation metrics of the proposed fraudulent website detection model.

Table 1. The proposed model performance using training set and cross validation.

Classification technique	Training			Cross validation		
	Accuracy	F-score	FPR	Accuracy	F-score	FPR
Naïve Bayes	100	98.33	1.62	95.34	95.02	4.87
Multinomial Naïve Bayes	99.61	99.55	1.04	99.41	99.11	1.64
Bernoulli	94.94	94.07	6.39	91.27	90.94	11.32
Support Vector Machine	90.27	89.67	12.52	88.37	87.38	14.01
Stochastic Gradient Descent	100	99.42	0.58	96.51	96.46	4.18
Logistic Regression	100	100	0	98.83	98.32	2.55
Voted Classifiers	98.64	97.53	2.54	97.67	97.25	3.49

Figure 3 shows the list of 20 discriminative words which appear in fraudulent website with relatively high frequency. Words like “user” “fast” “safe” “payment” “hour” are the most discriminative terms in fraudulent website detection. These words are usually appearing in phrases which aimed to lure the victims.

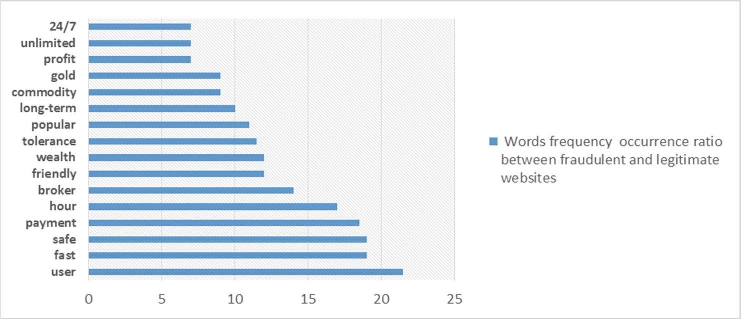

Fig. 3. Most discriminative words.

Figure 4 shows the ROC (Receiver Operating Characteristic) curves of the proposed fraudulent website detection model across different classifiers including Naïve Bayes, Multinomial Naïve Bayes, Bernoulli, SVM, Stochastic Gradient Descent and Logistic Regression. In this research eight different thresholds have been used to generate the ROC curve and measure the AUC (Area Under Curve). Figure 4 indicates that apart from SVM classifier, all other classifiers have relatively similar performance. SVM has relatively unsatisfactory performance in classification of bag of word features. Meanwhile Naïve Bayes and Logistic Regression classifiers marginally outperform other classifiers. The AUC (Area Under Curve) of each ROC curve is shown in Table 2. One major drawback of the existing solution is lack of scalability. Existing techniques might initially generate eye-catching results however these techniques are unable to keep pace with viral growth of fraudulent website. The proposed model deployed in the Big Data platform to ensure that it can keep up with vast amount and viral growth of fraudulent websites. Another major drawback of the existing fraudulent website detection models is their limited capabilities in data acquisition. Recent web programming languages and technologies made the web crawling and scrapping process more difficult than ever before. This study tackled this issue by means of efficient crawling engine which is able to deliver desired textual information regardless of the employed web technologies or deliberate limitations imposed by the websites. Using ensemble of classifiers, this study also managed to maintain its robustness against diversity and variety of the fraudulent website. One classification model is not able to perfectly generalized wide variety of fraudulent website.

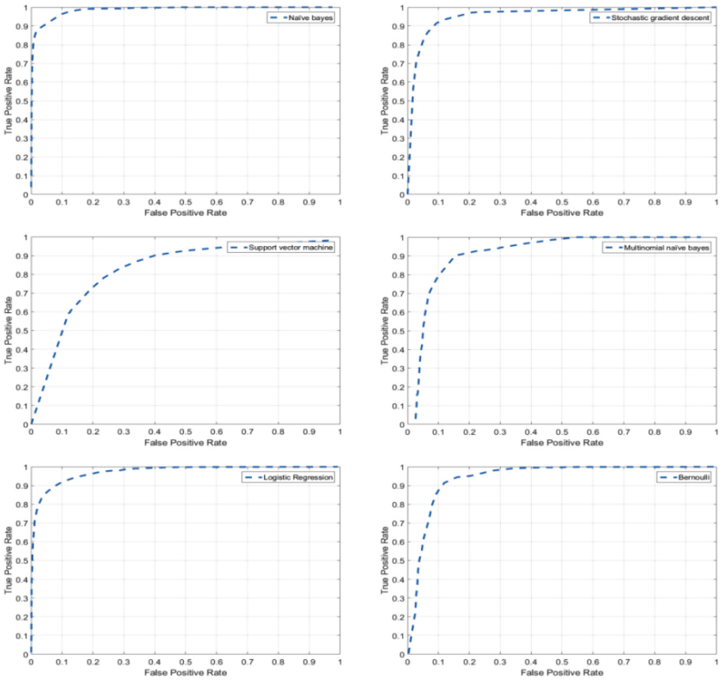

Fig. 4. ROC Curves of the proposed fraudulent website detection model across different classifiers.

Table 2. Area Under Curve (AUC) of the proposed fraudulent website detection model across different classifiers

Classifier	Area Under Curve (AUC)
Naïve Bayes	0.984
Multinomial Naïve Bayes	0.932
Bernoulli	0.958
Support Vector Machine	0.892
Stochastic Gradient Descent	0.971
Logistic Regression	0.979

5 Conclusion

Websites has turned into a platform for fraudsters to prey on victims and propagate fraud and cybercrime. Despite the efforts of researchers, majority of the measures are unable to keep pace with viral growth and diversity of fraudulent websites. This study attempted to address this issue by proposing a content based fraudulent website detection model which utilizes textual contents of web, natural language processing techniques and supervised classification techniques to counter fraudulent websites. Experiment results showed that the proposed fraudulent website detection model with cross validated accuracy of 97.67% and FPR of 3.49% achieved satisfactory results and served the aim of this research.

Acknowledgement. This work is supported by the Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under Fundamental Research Grant (FRGS) VOT R.J130000.7828.4F809.

References

1. Perner, P.: Advances in Data Mining: Applications and Theoretical Aspects. In: Proceedings of 10th Industrial Conference, ICDM 2010, 12–14 July 2010, vol. 6171. Springer, Heidelberg (2010)
2. Abbasi, A., Chen, H.: A comparison of tools for detecting fake websites. Computer **42**(10), 78–86 (2009)
3. Abbasi, A., Chen, H.: Detecting fake escrow websites using rich fraud cues and kernel based methods. In: Annual Workshop on Information Technologies and Systems, pp. 1–6 (2007)
4. Mohammad, R.M., Thabtah, F., McCluskey, L.: Tutorial and critical analysis of phishing websites methods. Sci. Rev. **17**, 1–24 (2015)
5. Phua, C., Lee, V., Smith, K., Gayler, R.: A comprehensive survey of data mining-based fraud detection research. In: 2010 International Conference on Intelligent Computation Technology and Automation, ICICTA 2010, vol. 1, pp. 50–53 (2010)
6. Le, A. and Markopoulou, A.: PhishDef: url names say it all. In: INFOCOM Proceedings IEEE, pp. 191–195 (2010)
7. Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., Nunamaker Jr., J.F.: Detecting fake websites: the contribution of statistical learning theory. MIS Q. **34**(3), 435–461 (2010)
8. Martines-romo, J., Araujo, L.: Web spam identification through language model analysis. In: Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web, pp. 21–28 (2009)
9. Urvoy, T., Lavergne, T., Filoche, P.: Tracking web spam with hidden style similarity. In: AIRWeb, pp. 25–31 (2006)
10. Ntoulas, A., Hall., B., Najork, M., Manasse, M., Fetterly, D.: Detecting spam web pages through content analysis. In: Proceedings of 15th International Conference on World Wide Web, pp. 83–92 (2006)
11. Shen, G., Gao, B. Liu, T. Y., Feng, G., Song, S., Li, H.: Detecting link spam using temporal information. In: Proceedings of IEEE International Conference on Data Mining, ICDM, vol. 49, pp. 1049–1053 (2006)
12. Becchetti, L., Donato, D., Baeza-yates, R., Leonardi, S.: Link analysis for web spam detection. ACM Trans. Web. **2**(1), 1–42 (2007)
13. Drost, I., Scheffer, T.: Thwarting the nigritude ultramarine: learning to identify link spam. In: European Conference on Machine Learning. LNCS(LNAI), vol. 3720, pp. 96–107 (2005)
14. Abbasi, A.: Detecting fake medical web sites using recursive trust labeling. ACM Trans. Inf. Syst. **30**(4), 1–22 (2012)
15. Liu, W., Deng, X., Huang, G., Fu, A.Y.: An antiphishing strategy based on visual similarity assessment. IEEE Internet Comput. **10**(2), 58–65 (2006)
16. Chou, N., Ledesma, R., Teraguchi, Y. Boneh, D., Mitchell, J.C., Ca, S.: Client-side defense against web-based identity theft. In: NDSS, pp. 1–16 (2004)

17. Abbasi, A., Zhang, Z., Chen., H.: A Statistical Learning Based System for Fake Website Detection, no. 4, pp. 3–4 (2008)
18. Ignatow, G., Mihalcea, R.: Text Mining: A Guidebook for the Social Sciences. Sage Publication, Los Angeles (2016)
19. Brill, E.: A simple rule-based part of speech tagger. In: Proceedings of the workshop on Speech and Natural Language 1992, pp. 112–116 (1992)