A Classifier to Detect Profit and Non Profit Websites Upon Textual Metrics for Security Purposes

Yahya Tashtoush[1,*], Dirar Darweesh[1], Omar Darwish[2], Belal Alsinglawi[3] & Rasha Obeidat[1]

[1]Department of Computer Science, Jordan University of Science and Technology,
Al Ramtha, Irbid, 22110, Jordan
[2]Information Security & Applied Computing, Eastern Michigan University,
201 Sill Hal, Ypsilanti, MI 48197, USA
[3] School of Computer Data and Mathematical Sciences, Western Sydney University,
Rydalmere, Sydney, 2751, Australia
*E-mail: yahya-t@just.edu.jo

Abstract. Currently, most organizations have a defense system to protect their digital communication network against cyberattacks. However, these defense systems deal with all network traffic regardless if it is from profit or non-profit websites. This leads to enforcing more security policies, which negatively affects network speed. Since most dangerous cyberattacks are aimed at commercial websites, because they contain more critical data such as credit card numbers, it is better to set up the defense system priorities towards actual attacks that come from profit websites. This study evaluated the effect of textual website metrics in determining the type of website as profit or nonprofit for security purposes. Classifiers were built to predict the type of website as profit or non-profit by applying machine learning techniques on a dataset. The corpus used for this research included profit and non-profit websites. Both traditional and deep machine learning techniques were applied. The results showed that J48 performed best in terms of accuracy according to its outcomes in all cases. The newly built models can be a significant tool for defense systems of organizations, as they will help them to implement the necessary security policies associated with attacks that come from both profit and non-profit websites. This will have a positive impact on the security and efficiency of the network.

Keywords: *classifier; cyber-attacks; defense system; network traffic; nonprofit; profit; security polices; textual metrics; website.*

1 Introduction

Nowadays, people commonly use websites for communicating with different types of institutions [1]. Websites are significant for both users and owners to achieve their goals. Websites can differ from each other according to the type of services they provide. Sites that are developed for profit institutions who are interested in providing financial services are called profit websites. The American shopping website Ebay.com is an example of a profit website [2]. The other type

Received September 15th, 2021, 1st Revision December 23rd, 2021, 2nd Revision February 20th, 2022, Accepted for publication April 6th, 2022.
Copyright © 2022 Published by IRCS-ITB, ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2022.16.1.6

is called nonprofit websites, which are developed for nonprofit institutions and provide public informational services for users. Yarmouk University's website (https://www.yu.edu.jo) is an example of a nonprofit website.

Every website on the Internet is vulnerable to security attacks. The threats range from human mistakes to advanced attacks by organized cyber criminals. According to the Investigations Report of Data breaking by Verizon, the main drive for cyber attackers is financial. So, whether you run an eCommerce project or a simple business website, the possibility of an attack is always there [3]. This is because businesses usually save the data of customers' bank accounts and credit cards, mailing addresses, email addresses, usernames and passwords. Cyber security attackers utilize these data for gaining money by credit card fraud or the use of consumers' private information for personality theft or fraud [4]. Therefore, website security is a major issue for business and profit companies [5].

An organization with a defense system to protect against cyberattacks has to be aware of cyber security threats that come from profit websites more than from nonprofit websites. Business websites are more vulnerable to security attacks because they keep more sensitive data such as credit card numbers.

Each cyberattack on an organization's network has its own characteristics and with the wide range of different kinds of attacks going around, it may seem impossible to protect your network against all of them [3]. Instead, you can direct your defense system towards actual threats and dangerous attacks, which come from high-traffic websites, such as profit websites. This would enhance the level of security and have a positive effect on the performance of the network. When the defense system only has to deal with specific attacks compared to when having to deal with all of them, minimizes network delay, because it enforces only the necessary security policies on the network, which is reflected in the speed of the network.

For this reason, our research built a classifier that can detect profit from non-profit websites. This can be very useful for companies to set the proper security level for their defense systems without network overhead, which can help these institutions to increase the security and efficiency of their network.

2 Related Work

Many research papers developed web page classification using mining techniques, but none of them studied classification of web pages into profit and nonprofit based on textual analysis for security purposes, like we have done in our research project. Some of these previous studies are discussed below.

Babapour and Roostaee [6] proposed an approach to address the challenge of classifying significant web content (short-term web content and long-term web content), where such a method could enhance search engine performance. They classified web pages into these two types by using machine learning techniques. They used natural language processing in addition to text mining methods for pre-processing the data, and then applied the machine learning techniques for classifying web pages. Qazi and Goudar [7] proposed a feasible solution to the problems associated with web page classification using a method called Ontology-based Term Weighting. Their approach depends on constructing a domain ontology and then choosing elements that can enhance the classification process. They conducted an experiment to evaluate their approach and came up with promising results.

Sun, *et al.* [8] implemented support vector machine (SVM) by grabbing web pages from different sites into different classes and taking both the content and context elements into consideration. Their classification approach was verified on a dataset called WebKB. This classification method produced better results than other classification approaches such as Foil-Pilfs on the same data set. The authors also demonstrated that including context elements (chiefly hyperlinks) enhances the efficiency of classification. The work by Hongjian and Yifei in [9] used an SVM classifier on a sample dataset and then applied article swarm optimization (PSO) to optimize the parameters. During the testing process of their new method about 100,000 web pages were gathered. F-measure was used to evaluate the empirical results, which indicated that their approach outperformed the SVM technique.

Chun, *et al.* [10] proposed a technique for extracting the information content from news web pages based on density characteristics. In their approach, HTML documents are divided into several blocks. The next step is to compute a value for each document according to particular density characteristics. Finally, the C4.5 data mining technique is applied to generate a classification of the documents blocks. This approach makes extracting information from news web pages simpler. Empirical tests indicated that this technique is easy and efficient for extracting news information. The work by Yazdani, *et al.* in [11] implemented naïve-Bayes models for each category of web pages and expanded the Hidden Markov model that was applied. A group of websites was used to compute the parameters of the models. Websites were formed in a tree pattern to classify them, and an algorithm called Viterbi was modified to estimate the possibility of producing these structures by each model.

Fiol-Roig, *et al.* [12] used a classifier (decision tree) that can classify web pages automatically, supporting search engines in retrieving the desired information. In their paper, they also tested the probability of using an automatic classifier.

Various classifiers using different mining methods that were exploited to build these classifiers were proposed. Ali, *et al.* in [13] presented the utilization of linear discriminant analysis (LDA), which is a general multivariate statistical data analysis method. It can enhance the classification processes of several classification models. LDA depends on the idea of gaining isolation among groups. It is usually utilized for reducing the dimensionality of datasets. Ali and Abdullah [14] introduced fast HP-PL as a novel parallel method for simplifying dimensionality reduction. It also enhances the accuracy of big data classification by utilization of the computational abilities of distributed-memory clusters. The method was implemented on Apache Spark. The authors also explain the importance of dimensionality reduction. This is a data mining technique that has become very popular and is an important step in many ML methods. Ali and Abdullah in [15] introduced a new parallel application of grid optimization using Spark Radoop to decrease large computation loads and facilitate the processing of big data.

3 Methodology

Our proposed research project consists of four phases, which are shown in Figure 1.

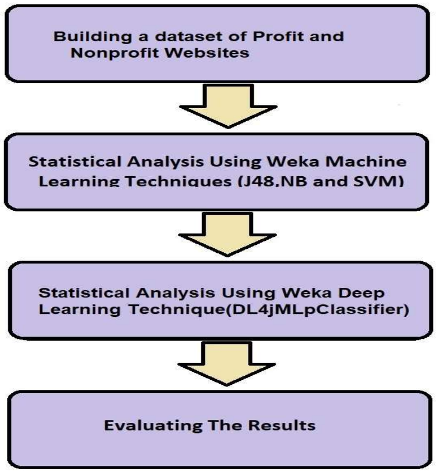

Building a dataset of Profit and Nonprofit Websites
Statistical Analysis Using Weka Machine Learning Techniques (J48,NB and SVM)
Statistical Analysis Using Weka Deep Learning Technique(DL4jMLpClassifier)
Evaluating The Results

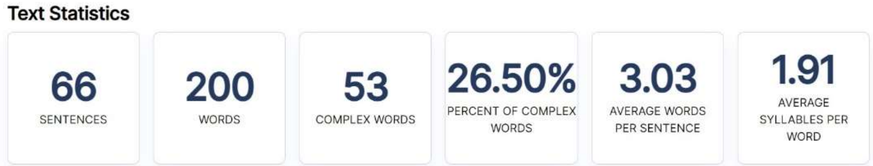
Figure 1 The research project work flow.

These research project phases are explained in detail in the following sections.

3.1 Building a Profit and Nonprofit Websites Dataset

This was the first step in our research project, where various websites were collected manually. These websites included profit and non-profit websites. Profit websites are marketing websites while non-profit websites are public informational websites from institutions such as universities, hospitals and ministries.

We used Readability Test Tool to extract the textual metrics of these collected websites. The readability test tool is an easy and rapid tool that can assess the readability of published texts [16]. Readability Test Tool computes the textual metrics of a web page, i.e., number of sentences, number of words, number of complex words, percentage of complex words, average number of words per sentence, and average number of syllables per word, where compound words are words with three or more syllables [16]. Figure 2 shows the textual metrics for the King Saud University website, as an example of using Readability Test Tool.



Text Statistics

66	200	53	26.50%	3.03	1.91
SENTENCES	WORDS	COMPLEX WORDS	PERCENT OF COMPLEX WORDS	AVERAGE WORDS PER SENTENCE	AVERAGE SYLLABLES PER WORD

Figure 2 Readability Test Tool results for the King Saud University home page.

We built our dataset using the MS-Excel 2010 database management system. It consisted of 237 rows. There were no missing values in our dataset. Its characteristics are clearly shown in Table 1, while Table 2 refers to a sample from the dataset.

Table 1 Dataset characteristics.

Attribute	Value
NoOfSent	Real
NoOfWord	Real
NoOfComWord	Real
ComWord	Real
AvgWordSent	Real
AvgSyllWord	Real
ProfitOrNonProfit	Profit, NonProfit
WebsiteType	University, Hospital, Ministry,Business

Table 2 Sample of dataset.

NoOfSent	NoOfWord	NoOfComWord	ComWord	AvgWordSent	AvgSyllWord	Profit or NonProfit	Website Type

113	751	155	20.64	6.75	1.68	Profit	Business
117	1101	206	18.71	9.48	1.81	Profit	Business
140	423	75	17.73	4.61	1.78	NonProfit	University
385	2202	447	20.3	9.26	1.8	NonProfit	University
2	23	2	8.7	11.5	1.48	NonProfit	Hospital
98	842	191	22.68	9.3	1.79	NonProfit	Ministry

3.2 Statistical Analysis Using Weka Machine Learning Techniques (J48, NB and SVM)

We applied the machine learning techniques in Weka tool to our dataset to generate several patterns and rules. These data mining techniques were J48 decision tree, Naïve Bayes (NB) and Support Vector Machine (SVM) techniques.

3.3 Statistical Analysis Using Weka Deep Learning Technique (DL4jMLpClassifier)

In this research stage, we applied a Weka deep learning technique called DL4jMLpClassifier to our dataset to obtain several patterns and rules.

4 Results and Evaluation

We applied different machine learning techniques to our data set for generating patterns and rules. These data mining techniques were J48 decision tree, Naïve Bayes and Support Vector Machine. Each of these machine learning techniques was applied to different training datasets. They involved 2, 5 and 10 folds in addition to the 66% training datasets. On the other hand, we also applied a Weka deep learning technique called DL4jMLpClassifier to our dataset in order to measure its accuracy in classification.

4.1 Results When the Class Label is 'ProfitOrNonProfit'

Figure 3 shows the J48 decision tree when the class label values were 'Profit' and 'NonProfit'. In Figure 3, it can be seen that the websites were classified either as Profit or NonProfit according to two metrics, i.e., the percentage of complex words and the average syllables per word.

Tables 3, 4 and 5 shows the results for J48 decision tree, Naïve Bayes, and Support Vector Machine, respectively, when the class label values were Profit and NonProfit.

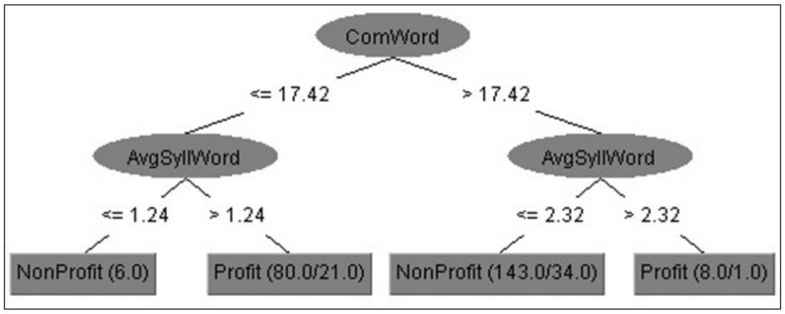


Figure 3 J48 decision tree (Profit and NonProfit).

Table 3 J48 decision tree classifier (profit and nonprofit).

Training Dataset	Correct	Percentage	Profit Precision	NonProfit Precision	Mean error	Profit F-Measure	NonProfit F-Measure
66% training	60	74.0741	0.526	0.806	0.395	0.488	0.826
2 folds	168	70.8861	0.725	0.702	0.3778	0.592	0.774
5 folds	162	68.3544	0.654	0.699	0.3999	0.586	0.744
10 folds	159	67.0886	0.641	0.686	0.4118	0.562	0.736

As can be seen in Table 3, the best result for J48 was achieved in the 66% training case, which had the highest number of instances that were classified correctly.

Table 4 Naive-base classifier (profit and nonprofit).

Training Dataset	Correct	Percentage	Profit Precision	NonProfit Precision	Mean error	Profit F-Measure	NonProfit F-Measure
66% training	58	71.6049	0.476	0.8	0.3356	0.465	0.807
2 folds	146	61.6034	0.574	0.631	0.3997	0.435	0.709
5 folds	142	59.9156	0.544	0.617	0.3979	0.395	0.7
10 folds	143	60.3376	0.552	0.62	0.397	0.405	0.703

In the case of NB, the best result was achieved in the 66% training case, which had the highest number of instances that were classified correctly, so this is the best choice, as can be seen in Table 4.

Table 5 Support vector machine classifier (profit and nonprofit).

Training Dataset	Correct	Percentage	Profit Precision	NonProfit Precision	Mean error	Profit F-Measure	NonProfit F-Measure
66% training	53	65.4321	0.412	0.83	0.3457	0.5	0.736
2 folds	145	61.1814	0.833	0.6	0.3882	0.179	0.746
5 folds	142	59.9156	0.647	0.595	0.4008	0.188	0.734
10 folds	145	61.1814	0.786	0.601	0.3882	0.193	0.744

Table 5 shows that the best choice for SVM was the 66% training case, which had the highest number of instances that were classified correctly. Overall, J48 was shown to be the best classifier according to its outcomes in all cases.

4.2 Results When the Class Label is 'WebsiteType'

Figure 4 in the Appendix shows the J48 decision tree when the values of the class label were University, Hospital, Ministry, and Business. The outcomes in Figure 4 were probably influenced by the number of each website type. The outcomes indicate that the most significant metric that distinguishes business websites from other types of websites is the Profit or Nonprofit metric. Figure 4 shows that a website can be classified as Business website only when the value of the profitOrnonprofit metric is equal to 'profit'.

Tables 6, 7 and 8 shows the outcomes of J48 decision tree, NB, and SVM, respectively, when the values of the class label were University, Hospital, Ministry, and Business.

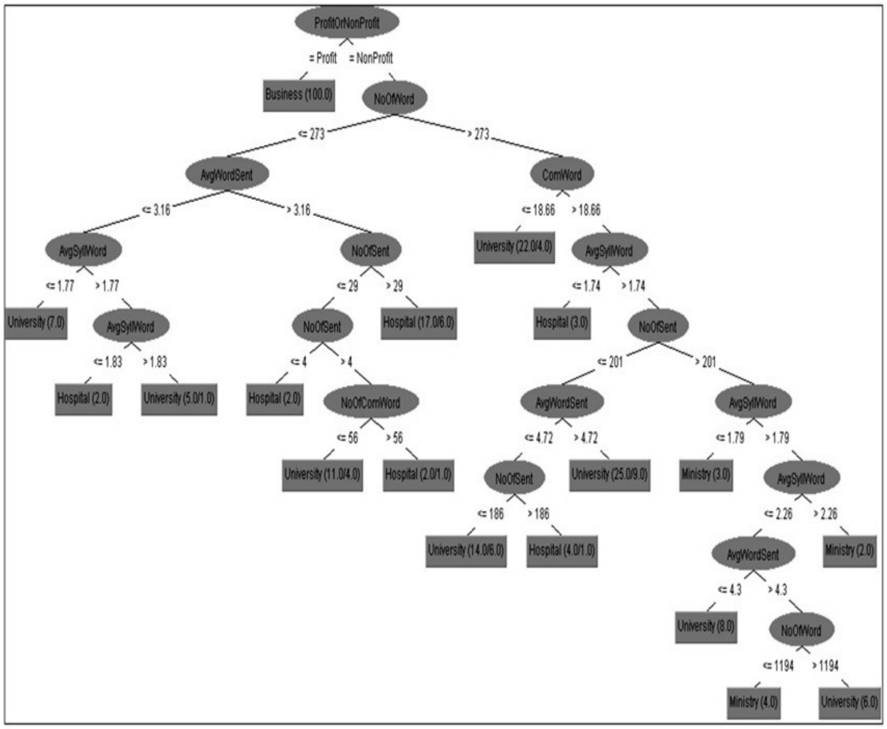


Figure 4 J48 decision tree (University, Hospital, Ministry, and Business).

Table 6 J48 decision tree classifier (university, hospital, ministry, and business).

Training Dataset	Correct	Correct Percentage	Incorrect	Incorrect Percentage	Mean error	F-Measure
66% training	51	62.963	30	37.037	0.2161	0.58
2 folds	174	73.4177	63	26.5823	0.1663	0.68
5 folds	173	72.9958	64	27.0042	0.1686	0.678
10 folds	172	72.5738	65	27.4262	0.166	0.675

As can be seen in Table 6, the best result for J48 was in the 2 folds case, which had the highest number of instances that were classified correctly.

Table 7 Naive-base classifier (university, hospital, ministry, and business).

Training Dataset	Correct	Correct Percentage	Incorrect	Incorrect Percentage	Mean error	F-Measure
66% training	51	62.963	30	37.037	0.233	0.59
2 folds	153	64.557	84	35.443	0.2181	0.646
5 folds	155	65.4008	82	34.5992	0.2073	0.63
10 folds	154	64.9789	83	35.0211	0.2045	0.637

The best result for NB was in the 5 folds case, which had the highest number of instances that were classified correctly, so this is the best choice as can be seen in Table 7.

Table 8 Support vector machine classifier (university, hospital, ministry, and business).

Training Dataset	Correct	Correct Percentage	Incorrect	Incorrect Percentage	Mean error	F-Measure
66% training	57	70.3704	24	29.6296	0.2881	0.593
2 folds	177	74.6835	60	25.3165	0.2802	0.656
5 folds	177	74.6835	60	25.3165	0.2799	0.656
10 folds	177	74.6835	60	25.3165	0.2795	0.656

Table 8 shows that the best results for SVM were obtained for 2, 5, 10 folds. Overall, SVM was shown to be the best classifier according to its outcomes in all cases.

4.3 Results when applying Weka Deep Learning Technique (DL4jMLpClassifier)

Table 9 shows the results of the DL4jMLpClassifier when the values of the class label were Profit and NonProfit. We applied this deep learning technique to three training datasets, i.e., 2, 5 and 10 folds, with different numbers of layers. The numbers of layers used were 2, 5, 7, and 9 layers. As shown in Table 9, DL4jMLpClassifier did not achieve high prediction accuracy compared with the traditional mining techniques. This was because the dataset used was not large, and deep learning requires large datasets to come up with good prediction results.

Table 9 DL4jMLpClassifier (Profit and NonProfit).

Training Dataset	2 Layers	5 Layers	7 Layers	9 Layers
2 folds	41.3502	57.8059	57.8059	57.8059
5 folds	41.3502	57.8059	57.8059	57.8059
10 folds	41.3502	57.8059	57.8059	57.8059

5 Conclusions and Future Work

As most cyberattacks target business websites, due to keeping more important data such as credit cards, it is highly necessary to set up priorities of the defense systems towards attacks from such websites. In this study, we built a classifier that can classify profit and non-profit websites according to some textual website metrics for security purposes. Deep neural networks, J48 decision tree, Naïve Bayes, and Support Vector Machine techniques were applied to a website dataset to create classifiers. The results indicated that websites were classified as profit or non-profit according to two primary metrics, i.e., the percentage of complex words and the average number of syllables per word. J48 performed the best in terms of accuracy according to its results in all cases. The new classifiers can assist cyber defense systems in applying the needed security policies and enhance the security and efficiency of the network. Future work includes building a classifier that can detect profit and non-profit websites according to website multimedia features.

References

[1] Gangeshwer, D.K., *E-Commerce or Internet Marketing: A Business Review from Indian Context*, International Journal of u-and e-Service, Science and Technology, **6**, pp.187-194, 2013.

[2] Ebay.com, https://www.ebay.com.au/ (7 Sept 2021).

[3] Svaiko, G., *The 10 Most Common Website Security Attacks*, https://www.tripwire.com/state-of-security/featured/most-common-website-security-attacks-and-how-to-protect-yourself/, (15 Dec 2021).

[4] Blog, I.S.B., *Common Cybersecurity Threats for E-Commerce Businesses*, https://www.insureon.com/blog/top-cybersecurity-threats-for-ecommerce-businesses, (15 Dec 2021).

[5] Johnson, N., *Why Website Security is Important for Your Business*, https://www.inmotionhosting.com/blog/why-website-security-is-important-for-your-business/, (7 Sept 2021).

[6] Babapour, S.M. & Roostaee, M., *Web Pages Classification: An Effective Approach Based on Text Mining Techniques*, IEEE 4th International

Conference on Knowledge-Based Engineering and Innovation (KBEI), 2017.

[7] Qazi, A. & Goudar, R.H., *An Ontology-based Term Weighting Technique for Web Document Categorization*, International Conference on Robotics and Smart Manufacturing, **133**, pp. 75-81, 2018.

[8] Sun, A., Lim, E.P. & Ng, W.K., *Web Classification Using Support Vector Machine*, Proceeding WIDM '02 Proceedings of the 4th International Workshop on Web Information and Data Management, pp. 96- 99, 2002.

[9] Hongjian, G. & Yifei, C., *Web Classification Algorithm Using Support Vector Machine and Particle Swarm Optimization*, IJACT, **4**(17), pp. 514 -520, 2012.

[10] Chun, Y., Yazhou, L. & Qiong, Q., *An Approach for News Web-Pages Content Extraction Using Densitometric Features*, Advances in Electric and Electronics Lecture Notes in Electrical Engineering, **155**, pp. 135-139, 2012.

[11] Yazdani, M., Eftekhar, M. & Abolhassani, H., *Tree-Based Method for Classifying Websites Using Extended Hidden Markov Models*, Advances in Knowledge Discovery and Data Mining, 13th Pacific-Asia Conference, pp.780-787, 2009.

[12] Fiol-Roig, G., Miró-Julià, M. & Herraiz, E., *Data Mining Techniques for Web Page Classification*, Highlights in Practical Applications of Agents and Multiagent Systems Advances in Intelligent and Soft Computing, **89**, pp. 61-68, 2011.

[13] Ali, A.H., Hussain, Z.F. & Abd, S.N., *Big Data Classification Efficiency Based on Linear Discriminant Analysis*, Iraqi Journal for Computer Science and Mathematics, pp. 2788-7421, September, 2020.

[14] Ali, A.H. & Abdullah, M.Z., *A Novel Approach for Big Data Classification Based on Hybrid Parallel Dimensionality Reduction Using Spark Cluster*, Computer Science, **20**(4), pp.413-431, December 2019.

[15] Ali, A.H. & Abdullah, M.Z., *A Parallel Grid Optimization of SVM Hyperparameter for Big Data Classification using Spark Radoop*, Journal of Modern Science, **6**(1), 3, March 2020.

[16] Reviews, W., Readability Test Tool, https://www.webpagefx.com/tools/read-able/, (7 Sept 2021).