International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume- 9, Issue- 6 (December 2019)
www.ijemr.net https://doi.org/10.31033/ijemr.9.6.18

Malicious-URL Detection using Logistic Regression Technique

Vanitha N[1] and Vinodhini V[2]
[1]Assistant Professor, Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, INDIA
[2]Associate Professor, Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, INDIA

[1]Corresponding Author: vanitha@drngpasc.ac.in

ABSTRACT

Over the last few years, the Web has seen a massive growth in the number and kinds of web services. Web facilities such as online banking, gaming, and social networking have promptly evolved as has the faith upon them by people to perform daily tasks. As a result, a large amount of information is uploaded on a daily to the Web. As these web services drive new opportunities for people to interact, they also create new opportunities for criminals. URLs are launch pads for any web attacks such that any malicious intention user can steal the identity of the legal person by sending the malicious URL. Malicious URLs are a keystone of Internet illegitimate activities. The dangers of these sites have created a mandates for defences that protect end-users from visiting them. The proposed approach is that classifies URLs automatically by using Machine-Learning algorithm called logistic regression that is used to binary classification. The classifiers achieves 97% accuracy by learning phishing URLs.

Keywords— URL, Logistic Regression, Machine Learning, Data

I. INTRODUCTION

Phishing websites are being employed to steal personal information, such as credit cards and passwords, and to implement drive-by downloads. Phishing is popular among muggers since it is easier to trick someone. In most cases, such annoying activity engages network resources intended for other use into clicking a malicious link which seems legitimate than trying to break through a computer's defense systems**.** In most cases, such annoying activity engages network properties intended for other uses, and nearly always threatens the security of the network and/or its data. Properly designing and deploying a Phishing URL will help block the intruders. phishing domain (or Fraudulent Domain) characteristics, the features that discriminate them from appropriate domains, why it is important to detect these domains, and how they can be detected using machine learning techniques.

Background Study

This section discusses related methodologies used by researchers who have tried to solve the problem of phishing URL detection and classification.

The authors Mohammed Nazim Feroz and, Susan Mengel[3] has describes an approach that classifies URLs automatically based on their lexical and host-based features. These methods are able to learn highly analytical models by extracting and automatically Mahout is established for such scalable machine learning problems, and online learning is considered over batch learning. The classifier achieves 93-95% accuracy by detecting a large number of phishing hosts, while maintaining a modest false positive rate.

Justin Ma, Lawrence K. Saul, Stefan Savag and, Geoffrey M. Voelker[4] describes an approach to this problem based on automated URL classification, using statistical methods to discover the tell-tale lexical and host-based properties of malicious Web site URLs. These methods are able to learn highly analytical models by extracting and repeatedly examining tens of thousands of features potentially indicative of suspicious URLs. The resulting classifiers obtain 91-94% accuracy, detecting large numbers of malicious Web sites from their URLs, with only modest false positives.

Frank Vanhoenshoven, Gonzalo N apoles, Rafael Falcot, Koen Vanhoof and Mario K¨oppent Universiteit Hasselt Campus Diepenbeek [1]determines online learning approaches for detecting malicious Web sites (those involved in criminal scams) using lexical and host-based features of the related URLs. We show that this application is mostly suitable for online algorithms as the size of the training data is larger which can be efficiently processed in batch also the distribution of features that typify malicious URLs is changing unceasingly.

II. PROPOSED METHOD

To ripen a defined manners from the data-sets, the model is to be sketched out like obliquely identify the data from which it has to be practised. The pillar of this model is data-sets and hence it should be sufficient and perfect data for good as well as bad URLs existing in the data for the model to be trained upon. A list of URLs that have been classified as either malicious or benevolent and characterize each URL via a set of attributes such as number of dots presents in URL, distance of the URL, token-based diagrams such as google.com. To train a model, binary classification technique which is also called as binary regression technique is used in a model.

Advantage of proposed method

- The proposed method acquires maximum learning accuracy comparing to other machine learning algorithms.

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume- 9, Issue- 6 (December 2019)
https://doi.org/10.31033/ijemr.9.6.18

- It consumes less time to learning phishing URLs.

III. UNIFORM RESOURCE LOCATOR (URL)

A URL is a exclusive identifier used to locate a resource on the internet. It is also denoted to as a web address. URLs consist of multiple parts -- including a protocol and domain name -- that tell a web browser how and where to recover a resource. End operators use URLs by typing them directly into the address bar of a browser or by ticking a hyperlink found on a webpage, bookmark list, in an email or from additional application. A URL is the most collective type of Uniform Resource Identifier (URI). URIs are strings of typescripts used to identify a source over a network. URLs are vital to traversing the internet.

URL Structure

The URL encompasses the name of the protocol required to access a resource, as well as a resource name. The first portion of a URL identifies what protocol to use as the primary access medium. The second portion identifies the IP address or domain name -- and possibly subdomain -- where the resource is located.

After the domain, a URL can also specify:

- A path to a exact page or file within a domain;
- A network port to use to make the link.
- A request or search parameters used -- commonly found in URLs for search results.

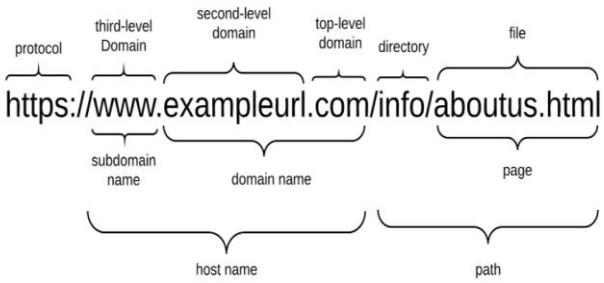

Fig 1: Structure of URL

Malicious URL

Within the gathering of cyber threats out there, mischievous websites play a critical role in today's attacks and scams.

Malicious URLs can be carried to users via email, text message, pop-ups or sheltered advertisements. The end effect can often be downloaded malware, spyware, ransom ware, compromised accounts. It should be obvious that being aware of what a Malicious URL is, and how it can do harm. Launch phishing movements meant to bargain your private information,

When we ticking a URL it directs us to phishing Sites and get you to install malware, viruses or Trojans, whether by transferring a file or as a drive-by-download that is provoked by something as simple as a mouse-over or other trick.

Example of Malicious URL

- timothycopus.aimoo.com
- cracks.vg/d1.php
- svisionline.de/ngfi/administrator/components/com_backup/classes/fx29id.com

IV. MACHINE-LEARNING APPROACH

Machine learning Methodology – This approach consists of two parts,
First chunk is Machine Learning model and the second chunk is Data-sets.

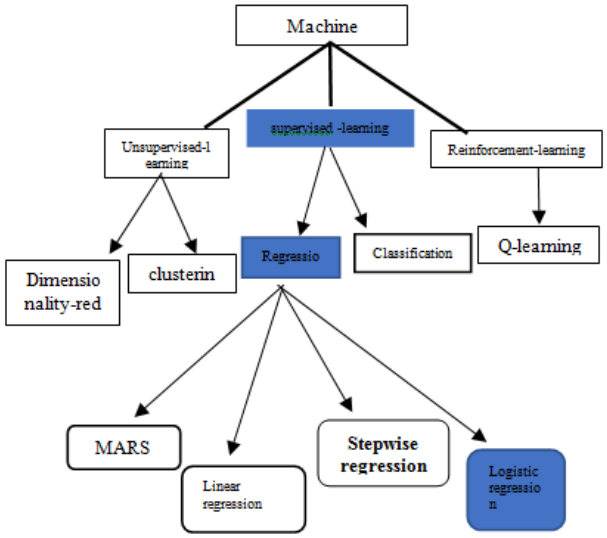

Fig 2: Classification of Machine Learning Algorithms

First Chunk- Machine learning

Machine learning is a subsection of artificial intelligence (AI) that offers systems the skill to mechanically learn and improve from experience without being explicitly programmed. Machine learning concentrates on the development of computer programs that can access data and use it learn for themselves.

The procedure of learning begins with data, such as examples, direct understanding, or instruction, in order to look for outlines in data and make better conclusions in the feature based on the examples that provide. The main aim is to permit the computers to learn automatically without human interference or assistance and regulate actions consequently.

Supervised learning

Supervised learning, in the background of artificial intelligence (AI) and machine learning, is a type of system in which both input and preferred output data are provided. Input and output data are labelled for classification to deliver a learning basis for future data processing. Supervised learning models have some benefits over the unsupervised approach, but they also have boundaries. The

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume- 9, Issue- 6 (December 2019)
https://doi.org/10.31033/ijemr.9.6.18

systems are more likely to make decisions that humans can relate to, for example, because humans have provided the basis for decisions. However, in the case of a retrieval-based method, supervised learning systems have distress dealing with new information.

Regression

Regression predictive modeling is the task of approaching a mapping function (f) from input variables (X) to a continuous output variable (y). A constant output variable is a real-value, such as an integer or floating point value

Stepwise regression

It is used when there is doubt about which of a set of analyst variables should be included in a regression model. It works by adding and/or removing separate variables from the model and detecting the resulting effect on its accuracy. Stepwise regression is no longer stared as a valid tool for dimensionality reduction because it yields unstable results that heavily over fit the training data.

Multivariate Adaptive Regression Splines (MARS)

It is a form of regression analysis. It is a non-parametric regression technique and can be seen as an extension of linear models that automatically models nonlinearities and communicate between variables.

Logistic Regression

The **logistic regression** technique includes dependent variable which can be signified in the binary (0 or 1, true or false, yes or no) values, means that the result could only be in either one form of two. For example, it can be applied when we need to find the probability of positive or fail event. Here, the same method is used with the additional sigmoid function, and the value of Y ranges from 0 to 1.Consider a model with two predictors, x1 and x2; these may be constant variables or indicator functions for binary variables (taking value 0 or 1). Fig 5 represents the comparison method [7].

1/(1+e^-x)

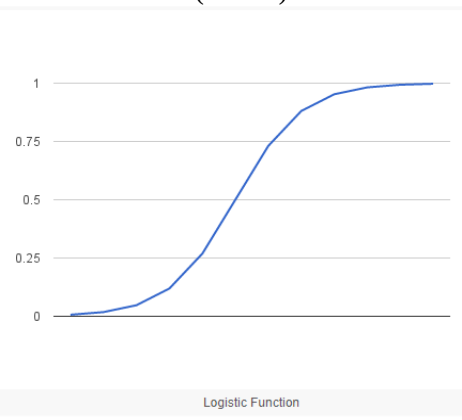

Fig 5: Logistic Function

Comparison of linear and Logistic Regression

Linear and Logistic regression are the furthermost basic form of regression which are usually used. The crucial difference between these two is that Logistic regression is used when the dependent variable is binary in nature. In difference, Linear regression is used when the dependent variable is continuous and nature of the regression line is linear.

Regression is a method is used to predict the value of a response (dependent) variables, from one or more predictor variables, where the variable is numeric. There are several forms of regression such as linear, multiple, logistic, polynomial, non-parametric, etc.

V. WORKING METHODOLOGY

System Flow

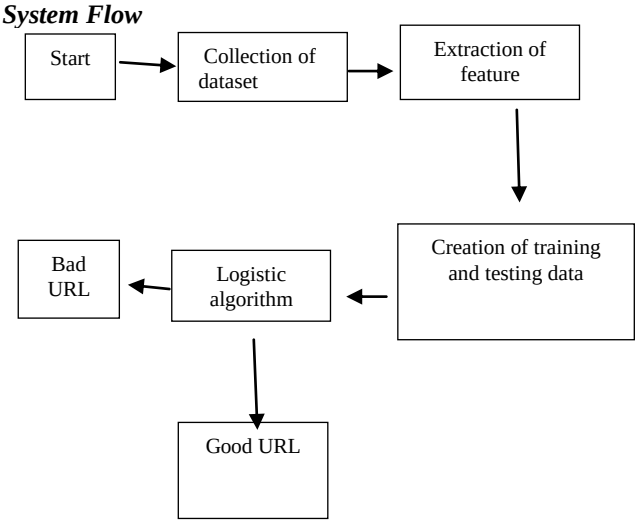

Fig4: system overflow

Second Chunk -Data Sets

The training data set in Machine Learning is the genuine dataset used to train the model for performing various actions. This is the actual data the current development process models learn with several API and algorithm to train the machine to work automatically.

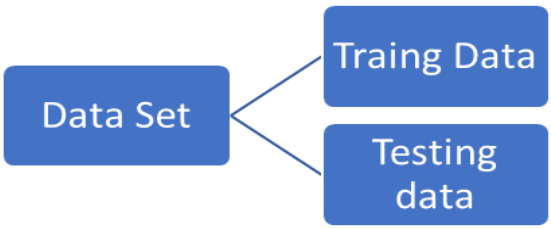

Fig6: Dataset Classification

Training Dataset

There are two types of data sets – Training, and Test that are used at several stage of development. Training dataset is the leading of two of them, while test data functions as closure of approval and you don't need to use till the end of the development.

Test Dataset

This is the data typically used to provide an balanced evaluation of the final that are completed and fit

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume- 9, Issue- 6 (December 2019)
https://doi.org/10.31033/ijemr.9.6.18

on the training dataset. Essentially, such data is used for testing the model whether it is responding or working properly or not.

- All of URL in our dataset are labelled
- Data sets are collected from

https://github.com/VAD3R-95/Malicious-URL-Detection/blob/master/data_URL.csv yahoo-phish tank

Extraction of Feature

In machine learning, a feature is an separate assessable property or characteristic of a phenomenon being detected. Picking informative, perceptive and independent features is a vital step for effective algorithms in pattern recognition, classification and regression. When the input data to an algorithm is too huge to be processed and it is suspected to be redundant. then it can be converted into a reduced set of features The selected features are expected to contain the appropriate information from the input data, so that the desired task can be performed by using this reduced demonstration instead of the complete initial data. Since the URLs are in our dataset are different from our normal text documents so we have to use text feature extraction method for construct a feature vector. Fig 7 shows the feature factorizing methods [8].

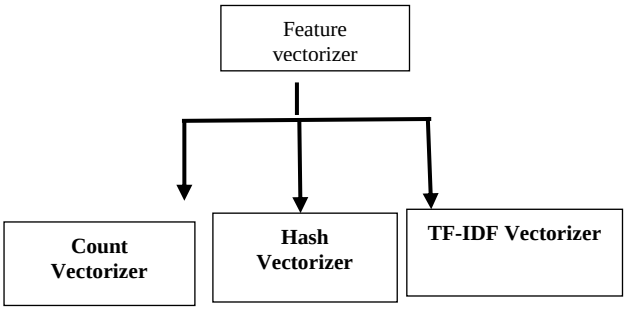

Fig 7: Feature Vectorizing Methods

Count Vectorizer

The most straightforward one, it counts the number of times a token shows up in the document and uses this value as its weight.

Hash Vectorizer

This one is measured to be as memory efficient as possible. In its place of storing the tokens as strings, the vectorizer applies the hashing trick to encode them as numerical indexes. The problem of this method is that once vectorized, the features' names can no longer be recovered.

TF-IDF Vectorizer

TF-IDF stands for "term frequency-inverse document frequency", means the weight allocated to each token not only depends on its frequency in a document but also how persistent that term is in the entire corpora.

Preparing Data

Subsequently the URLs are dissimilar from our typical text documents, we need to engrave our own *purification* method to get the appropriate data from raw URLs. To contrivance our distillation function in python to filter the URLs with following code as shown in trial code. This will give us the desired URL data-set values to sequence the model and test it. The data-set will partake two pillar, one is for *URLs* and other is for *labels*. Here we have proceeded with the use of **Tf-idf** machine learning text feature extraction approach from the python module of sk-learn.

Feature Vector Construction

http:// www. Bfuduuioolfb .mob /ws/ ebayisapi .dll

Fig:8 Vector Construction

Features Considered

- Blacklist Queries
- Lexical Features

Blacklist

- List of known malicious sites from yahoo phish tank, google crawlers.
- List of malicious URLs from various domain Providers like SORBS, URIBL, SURBL, Spamhaus.

Lexical Features

- Tokens in URL hostname + path
- Length of URL
- Entropy of the domain name

Reading Data

It is essential to recite the data-sets into data frames and matrix, which can be presumed by the Vectorizer. After Vectorizer data are arranged and distributed onto the *term-frequency and inverse document frequency*, which is called as text extraction approach. **Pandas** component in python is used for the task to be implemented.

Splitting Data

The data we use is typically split into training data and test data. The training set covers a known output and the model learns on this data in order to be universal to other data later on. The test dataset (or subset) is to test our model's prediction on this subset. In order to use the splitting method we have to import pandas library

- **training set**—a subset to train a model. (80%)
- **test set**—a subset to test the trained model. (20%)

Training Model

To train model call the logistic algorithm that is imported using sklearn model from python sci-kit library. (From sklearn.linear_model import Logistic Regression). It uses train data set for learning. After learning it prints score of trained model.

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume- 9, Issue- 6 (December 2019)
https://doi.org/10.31033/ijemr.9.6.18

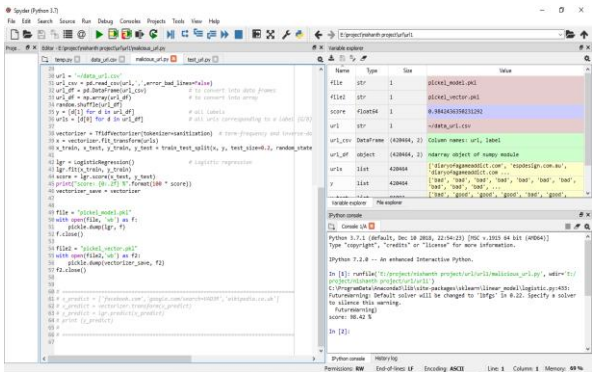

Fig 9. Training model

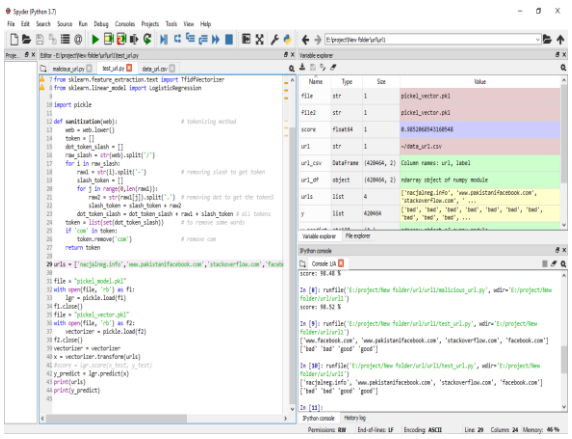

Fig 10: Testing Model

Testing Model

Pass the various URLs as inputs in to trained model. It predicts whether the URL is good or bad and returns the output as good/bad.

VI. EXPERIMENT AND RESULT DISCUSSION

The Model is to be sketched to detect malicious URLs by using machine learning methods. Machine learning model and datasets are the two dissimilar quantities of the process.

Table I. Learning Accuracy –random Splitting

Split ratio	Random forest	Naive Bayes	Logistic Regression
1:1	88.36	76.46	91.23
4:1	92.48	87.11	96.21
10:1	95.46	92.48	98.42

Since the URLs are always dissimilar from our usual script, to get appropriate data from URLs, own sanitization method is written using pandas package. Then the data in dataset are imported as data frames and arrays, for that machine-learning numpy package is used. The data can be understood by vectorizer which we prepared by using **Tf-idf** machine learning, this is the type of machine learning for text feature extraction method from the *python* module called *sklearn*. Then the logistic regression method is used to train and test our data.

The model is trained for multiple times with different data split ratio such as 1:1, 1:4, 10:1 and compared with previous models and the learning accuracy score is compared.

The Table I shows the comparison of Logistic Regression method with Other methods.

Training Model

The model is trained with various split ratios such as 1:1,4:1,10:1 and learning accuracy is noted. proposed method is compared with other algorithms such as naïve bays and random forest. The comparison of learning accuracy is shown in the below table.

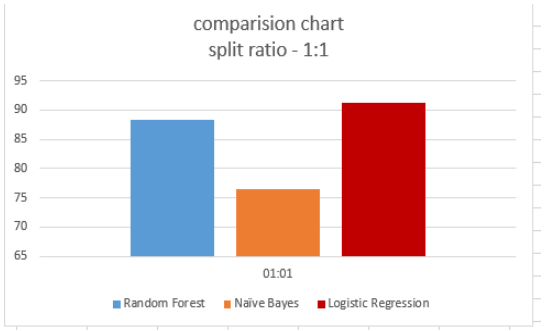

Fig. 11 : Comparison of Logistic Regression with Other methods with split ratio 1:1

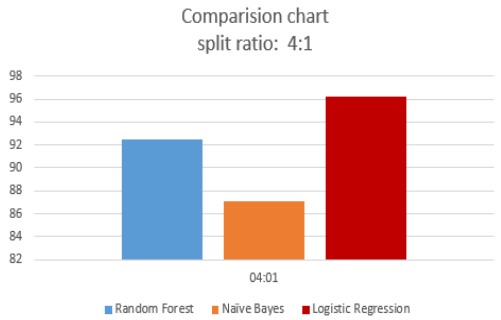

Fig. 12: Comparision of Logistic Regression with Other methods

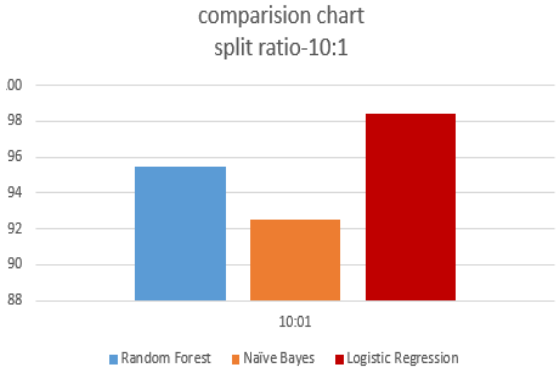

Fig. 13

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume- 9, Issue- 6 (December 2019)
https://doi.org/10.31033/ijemr.9.6.18

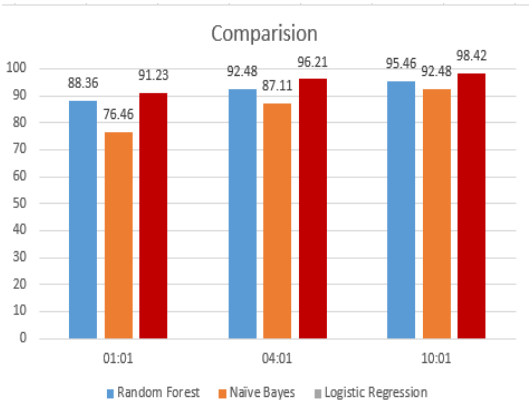

Fig. 14

VII. CONCLUSION AND FUTURE ENHANCEMENT

Malicious URL detection plays a serious role for many cyber security applications, and networking applications. The majority of computer attacks are launched by visiting a malicious webpage. A user can be tricked into voluntarily giving away private information on a phishing page or become target to a drive-by download resulting in a malware infection. In this approach we showed phishing URL detection by using machine learning algorithm called logistic regression, it obtains maximum learning accuracy comparing to other algorithms such as naïve bays, random forest. In future there is an idea to increase training and testing data and to find vary of accuracy, and can deploy as web content for all the network connected devices. In addition to that adding some more feature like host based (WHOIS) features makes our model more accurate.

REFERENCES

[1] Justin Ma, Saul L. K., Savage S., & Voelker G. M. (2011). Learning to detect malicious urls. *ACM Transactions on Intelligent Systems and Technology*, *3*(2), 1–24.

[2] Verma R. & Das A. (2017). Whats in a URL: Fast feature extraction and malicious URL detection. *In 3rd International Workshop on Security and Privacy Analytics*, pp. 55–63.

[3] Patil D. R. & Patil J. B. (2016). Malicious web pages detection using static analysis of URLs. *International Journal of Information Security and Cybercrime*, *5*(2), 57–70.

[4] Zuhair, H., Selamat, A., & Salleh, M. (2015). Selection of robust feature subsets for phish webpage prediction using maximum relevance and minimum redundancy criterion. *Journal of Theoretical and Applied Information Technology*, *81*(2), 188–205.

[5] Hajian Nezhad J, Vafaei Jahan M, Tayarani-N M, & Sadrnezhad Z. (2017). Analyzing new features of infected web content in detection of malicious web pages. *The ISC International Journal of Information Security*, *9*(2), 63–83.

[6] Mark Dredze, Koby Crammer, & Fernando Pereira. (2008). Confidence-weighted linear classification. *In 25th International Conference on Machine Learning (ICML)*, pp. 264–271.

[7] Hsu C. W. & Lin C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 415–425. Doi: 10.1109/72.991427.

[8] Crammer K., Dredze M., & Kulesza A. (2009). Multiclass confidence weighted algorithms. *In Conference on Empirical Methods in Natural Language Processing*, pp. 496–504.