
Phisher Fighter: Website Phishing Detection System Based on URL and Term Frequency-Inverse Document Frequency Values

E. Sri Vishva and D. Aju*

Vellore Institute of Technology, Vellore, Tamil Nadu, India

E-mail: daju@vit.ac.in

**Corresponding Author*

Received 28 May 2021; Accepted 14 October 2021;
Publication 28 October 2021

Abstract

Fundamentally, phishing is a common cybercrime that is indulged by the intruders or hackers on naive and credible individuals and make them to reveal their unique and sensitive information through fictitious websites. The primary intension of this kind of cybercrime is to gain access to the ad hominem or classified information from the recipients. The obtained data comprises of information that can very well utilized to recognize an individual. The purloined personal or sensitive information is commonly marketed in the online dark market and subsequently these information will be bought by the personal identity brigands. Depending upon the sensitivity and the importance of the stolen information, the price of a single piece of purloined information would vary from few dollars to thousands of dollars. Machine learning (ML) as well as Deep Learning (DL) are powerful methods to analyse and endeavour against these phishing attacks. A machine learning based phishing detection system is proposed to protect the website and users from such attacks. In order to optimize the results in a better way, the

Journal of Cyber Security and Mobility, Vol. 11-1, 83–104.

doi: 10.13052/jcsm2245-1439.1114

© 2021 River Publishers

TF-IDF (Term Frequency-Inverse Document Frequency) value of webpages is employed within the system. ML methods such as LR (Logistic Regression), RF (Random Forest), SVM (Support Vector Machine), NB (Naive Bayes) and SGD (Stochastic Gradient Descent) are applied for training and testing the obtained dataset. Henceforth, a robust phishing website detection system is developed with 90.68% accuracy.

Keywords: Phishing, machine learning, logistic regression, random forest, support vector machine, TF-IDF.

List of Notations and Abbreviations

ML (Machine learning), DL (Deep Learning), LR (Logistic Regression), RF (Random Forest), SVM (Support Vector Machine), NB (Naive Bayes), SGD (Stochastic Gradient Descent), PII (Personal Identity Information), ROC curves (Receiver Operating Characteristic), URL (Uniform Resource Locator), HTML (Hypertext Markup Language), GBDT (Gradient Boosting Decision Tree).

1 Introduction

Primarily, phishing is considered to be a felonious activity carried out by an intruder or a hacker through a well versed technical deceptions and social engineering. The sole purpose is to steal the sensitive PII (Personal Identity Information) as well as the pecuniary credentials of the users or customers. The social engineering strategy utilizes baiting as well as spoofed emails to trick the beneficiaries to disclose their financial data. The received mail claims to be from a legitimate user or business houses which in turn misleads the beneficiary to be deceived. Another interesting fact is the technical stratagem strategy which steals the sensitive credentials from the user's computer system by installing malicious software in it. The main intent of this scheme is to intercept the users' online credentials such as username as well as passwords.

Data breaches such as Privacy leaks, property thefts as well as identity thefts happens due to the well-known phishing method that will be used as computer network attacks. Throughout the year 2017, 29.4% of networked computers were considered to be attacked by at least one malware based web attack in accordance with the Kaspersky Laboratory statistics. And also, a number of 99,455,606 unique URLs were acknowledged as malevolent URLs

by the antivirus components. Additionally in the same year, out of all phishing detection, the financial phishing has grown to a larger extent from 47.5% to 54% approximately. Consequently, this kind of data breach has become one of the major security threat over the internet.

Day by day the cyber security strategies change and evolve in order to protect and safeguard the computer systems from intruders or hackers. On the other hand, as the security strategies evolve, intruders or hackers are evolving and unfolding themselves to develop more and more advanced security strategies to evade the respective protected systems. With the help of machine learning methods, the cybersecurity systems analyses the various data patterns and consequently learn and discover from the training and analysis. This respective training and analysis aids the cyber security systems to impede similar kind of web attacks. Another significant factor of machine learning scheme is to react to the altering behaviour of the web attacks in less amount of time.

To train up, interpret a machine learning model in an efficient way is the logistic regression classifier. In the feature space, this particular model does not make any suppositions about the dissemination of classes. It is observed and known that the logistic regression classifier is very quick in classifying the unfamiliar records. Concurrently, this classifier may lead to overfitting if the number of observance is below the number of selected features. Basically, random forest classifier is a predictive model that can be utilized as a regression model and also as a classification method. It provides variable importance since it aids in deciding the particular variable because of its positive influence. Random forest manages the variables very quick so as to make it suited for complex tasks. The primary shortcoming of this classifier is that it becomes too decelerate as well as ineffectual for real-time prognosis when it has significant number of trees.

When there is an obvious margin of division among the object classes, support vector machine works relatively fine. SVM is more efficient in high-dimensional spaces and also, it is observed that it is effectual when the number of dimensions are larger than the number of samples. When the number of features for each data point surpasses the number of training data samples, this classifier underachieves. The disadvantage of SVM classifier is that it does not work well for large number of datasets. Also, if the obtained dataset has more noise, this respective classifier does not perform well due to which the target classes overlaps.

Naive Bayes is an easy and powerful method for predictive modelling that builds upon strong suppositions of the covariates independence in

implementing Bayes theorem. It presumes independence among the predictor variables on the response as well as the Gaussian allocation. The respective distribution happens with mean and standard deviation that is calculated from the training dataset. Generally, for the classification problems, this classifier models are utilized as a substitute to decision trees. Stochastic Gradient Descent is an effective method for fitting linear classifiers as well as regressors in SVM and LR. One of the significance of SGD is its capability of handling large-scale learning where it is been employed to large-scale and sparse ML problems with respect to text classification. The advantage of this model is its efficacy and easier implementation. And the disadvantage is that it demands a number of hyper-parameters.

The contribution of this proposed work Phisher Fighter is as follows.

The overall proposed architecture itself is a unique framework that detects phishing in websites. The proposed framework is basically executed in two phases and in each phase of execution, different machine learning algorithms are employed on the given dataset to obtain the best algorithm. A score is calculated for the two machine learning algorithms that are considered best and effective based upon the trueness of URL and the probability of trust. If the score is 0, the respective website has encountered phishing and if it is 1, then the respective website has not encountered phishing. Also, based upon the literature survey conducted, it is understood that there is no similar research results where these set of algorithms have been utilized for phishing detection.

Section 1 gives a background about the website phishing and the significance of it. It also provides a brief idea about the various machine learning algorithms. Section 2 provides a detailed literature survey pertaining to the research work carried out with respect to website phishing. A total of 21 research articles are considered and presented for this particular work. Section 3 provides the overall architecture of the proposed system, Phisher Fighter. The architecture contains three phases of execution. The first phase is the URL analysis, Content analysis and Phishing detection. Section 4 provides the Phase-1 URL Analysis, which processes and analyses the websites based upon the web urls. Section 5 is the Phase-2 provides the Content Analysis which processes and analyses the websites based upon the web contents.

Section 6 is the phase-3 provides the phishing detection. where it really identifies whether the respective website is a phishing or non-phishing website. Section 7 details the dataset that is utilized in the proposed work to determine the websites are afflicted with phishing or not. Section 8 furnishes

the result analysis pertaining to the phishing detection of websites. The analysis shows the efficacy of the proposed methodology with respect to URL analysis and Content analysis. Also, a comparison of different different phishing detection techniques along with the proposed method is tabulated. Section 9 outlines the conclusion and future work of the proposed phishing methodology. Section 10 provides the references through which the research work is carried out.

2 Literature Survey

A trustworthy ensemble cataloguing system [1] is proposed to amalgamate the predicted outcome from various phishing detection classification methods. Also, a hierarchic clustering methodology has been utilized for the automatic cataloguing of phishing websites. Various classes of features such as headings, keywords and weblink information that is embodied in the webpages over the internet is extracted for the purpose of cataloguing. A phishing detection approach that catalogues the webpage security by checking its source code [2] is proposed. The security of the websites are evaluated by examining the phishing characteristics based on the W3C standards subsequently verifying the source code. On verifying the website source code, if any phishing character is found, the initial secure weightage is decreased. Finally the security level of the respective website is calculated based on the final weightage. The high level of web security denotes that the respective website is secured and the low level of web security indicates that the respective website is most likely to be a phishing site. An anti-phishing scheme that operates based on different cataloguing algorithms and the features based on natural language processing [3] is proposed. Various features of the webpage such as text similarity, font colour and size as well as the images from the webpage. It is observed that the text-based similarity methods are relatively faster and at the same time it fails to detect the attacks if the text is replaced with images. ML classifier methods in conjunction with the wrapper feature selection [4] technique is presented to detect the phishing websites accurately. It is noted that the RF classification method combined with the NLP features, provided the finest attainment of 97.98% accuracy for the criminal theft URL detection.

The discernment of important features [5] that distinguishes the legitimate as well as the phishing URLs are focussed. It is noted that the wrapper-based features took longer time as well as required more computation with few features. Here, the most number of significant features are rendered by the

utilization of the wrapper- based features. Additionally, it is observed that a selection of user-threatened features can be considered and incorporated to improvise the effectiveness of website acquisition. A novel cataloguing model [6] is proposed based upon the heuristic features of URL, the source-codes as well as the third-party aids. Primarily, all this features are extracted to surmount the drawbacks of the already present phishing protection methodologies. The suggested and implemented model is evaluated utilizing eight different machine learning algorithms. It is observed and noted that the suggested method exceeded in its performance when compared to all the other present methods. A novel classification method has been proposed [7] to surmount the corrupt activities of the anti-crime robberies. This classification is in accordance with the empirical features that are pulled out from the URL features, source-code features as well the third- party aid features. The developed model is been examined with eight different machine learning algorithms and it is observed that the RL classification method performed well with an efficacy of 99.31 percentage. It is observed that the sensitive identity theft schemes works better than the phishing scams as well as the user training solutions [8]. The reason being that it does not require a change in authentication platforms and does not depend on user's ability to identify the sensitive identity theft.

A phishing identification mechanism [9] that is based upon the visual similarity that mimic the identical victim website is proposed. It is observed and noted that the developed system was able to extract 224 unique web pages mimicked by 2262 phishing websites. A detection rate of over 80% is achieved while the false positive rate is observed at 17.5%. An intelligent system that detects phishing attacks [10] is presented. Different data mining methods are used to decide upon the categories of the websites. In order to evaluate the performance of the data mining methods, classification accuracy, ROC curves (Receiver Operating Characteristic, and F-measure were used. It is observed that the Random Forest has outperformed all other methods and provided an accuracy of 97.36%. A phishing detection approach named PhishZoo [11] is proposed that utilizes the profiles of trusted websites' appearances. The websites that are built with fuzzy hashing methodologies are used to detect phishing. Over 600 phishing websites that imitates 20 real websites were tested and it is observed that PhishZoo provided better accuracy with the blacklisting techniques. It also classifies new attacks and targeted attacks against smaller corporate intranets. Evaluation of websites whether it is legitimate or not. A contribution on improvising the phishing detection efficacy on websites is performed. For the purpose of improvisation,

a feature selection method is amalgamated along with an ensemble erudition methodology is employed. The developed method resulted with an accuracy rate of up to 95%.

A research that focusses upon evaluating whether or not a particular website is authentic is presented. In order to improvise the accuracy, a feature selection method that is amalgamated with an ensemble learning method [12] is employed. Pertaining to the phishing identification, the experimental results of the proposed method provides an accuracy of 95%, that is higher than the existing methodologies. Once the respective learning technique is deployed, the results provides an accuracy of 97% for identity theft. A feature selection and multiple ensemble learning model [13] is proposed to address the overfitting and at the same time improvising the prediction accuracy. With respect to the multiple learning techniques, the prediction is not biased against one single particular model. Rather, it will be in accordance with the majority of the predictions that is been already made. Therefore the predictions are made from each and every model that influences the final ensemble prediction. A method that evaluates correlation-based and wrapper feature selection methodology [14] is presented. In the process of experimentation, 177 initial feature sets from the real-world phishing data sets are utilized. Employing an efficient and effective feature selection results in important improvisation statistically in the accuracy of classification. Significant cataloguing accuracies are improvised among different models such as Random Forest as well as Logistic Regression.

A novel anti-phishing methodology [15] that employs a training intercession for detecting the phishing websites is proposed and evaluated. The proposed method aids the users in making the right decisions in differentiating the legitimate websites as well as the phishing websites. A novel technique that detects the phishing attacks [16] through the obtained website hyperlinks and proper analysis of it is been presented. The intended method segregates the features that pertains to website hyperlink features in to twelve different catalogues. These catalogued features are utilized to train the machine learning algorithms.

The proposed and developed method provides a correctness of 98.4% on the LR classification model. A heuristic model [17] that decides upon whether a particular website is authentic or a phishing site is presented. The proposed heuristic approach extracts twelve features and all the obtained features are trained up with support vector machine. Subsequently, the testing set fed into trained model to perform the respective testing. Comparatively, the proposed technique provides a higher accuracy rate with other existing models.

A stacking technique [18] to detect the phishing webpages through Uniform Resource Locator (URL) as well as Hypertext Markup Language (HTML) features are proposed and presented. Lightweight URL as well as HTML features are designed and subsequently these features establishes embedding of HTML string. By blending Gradient Boosting Decision Tree (GBDT), XGBoost as well as Light GBM in multiple layers, a stacking technique is been devised. It is observed that the proposed method was able to achieve an accuracy of 97.30% in detecting the phishing webpages. An improvised variant of the favicon oriented phishing attack detection [19] is proposed by introducing the domain name amplification features. When the websites does not have favicon, the additional feature are very much useful. An aggregate of 5000 each phishing websites are got from PhishTank as well as Alexa to check the effectiveness of the proposed technique. It is observed that the proposed technique was able to achieve and accuracy of 96.93%.

A study is aimed at cataloguing the news into different classes whereby the users can distinguish the most prevalent news groups at any point of time in any desired country. A new cataloguing technique [20] is proposed based upon TF-IDF as well as SVM. By utilizing the BBC datasets and also a five group of 20Newsgroup datasets, the proposed technique was able to achieve the cataloguing precision to be 97.84% with BBC dataset and 94.93% with the 20Newsgroup dataset. A phishing detection system is developed based upon the efficacy of the features of a scalable learning classifier [21]. Millions of pages are analyzed on a daily basis based upon the developed analyzer and it is observed that it correctly classifies 90% of the phishing pages. A natural language processing system [22] is developed that reads the plain English text and catalogues with respect to the conceptual topics and ontological construct. A phishing algorithm named PhishCatch [23] is developed to detect phishing websites. The developed algorithm is a heuristic algorithm that detects the phishing emails and subsequently alerts the user. The developed algorithm provides an accuracy of 99% in detecting the attacks. A design, execution and the examination of CANTINA [24] is presented. It is a novel content based technique to detect the phishing websites. Primarily, the developed CANTINA was based upon TF-IDF methodology for the retrieval. It is noticed that the experimental results shows that the developed technique was able to achieve an efficiency of 95% in the phishing website detection.

A novel methodology that works and protects against phishing attacks utilizing the autoupdated white lists of websites [25] I proposed. It is observed that the empirical results shows that the developed methodology provides

a detection rate of 86.02%. By analyzing the real internet protocols from the internet service provider, a detection technique is proposed based upon graph mining along with belief propagation [26] to protect the websites from phishing. An improvised phishing detection method named as 'Embedded Phishing Detection Browser' (EPDB) [27] that incorporates itself along with the browser architecture is introduced. The introduced method preserves the current user experience and at the same time improvises the security with a 99.36% of accuracy. A newfangled build of phishing that pertains to attacks, categories of attackers, weaknesses, targets, medium of attacks and its techniques [28] were proposed. Two different sets of website dataset is focussed and presented here to build an efficient phishing detection system. The two different datasets consists of 58,645 and 88,647 website datasets [29] that are classified as phishing and non-phishing.

3 Overall Architecture

The overall architecture can be divided into three parts such as URL analysis, content analysis and evaluating the final score of the web page by using the results obtained from URL analysis and Content analysis. Initially, the obtained dataset is segregated as training dataset and testing dataset. The obtained training dataset is pre-processed and the respective features are extracted. The extracted features are then utilized for training through the machine learning models. Three distinct machine learning models such as LR, RF and SVM are utilized to carry out the training of the dataset. Subsequently, the testing dataset is used to obtain the accuracy of all three models to find its efficiency. The machine learning model with the best accuracy is chosen as the selected model to perform classification so as to determine the trustworthiness of the respective URL.

Parallely, in the second phase, the overall dataset is segregated as training dataset and testing dataset. The testing dataset is then pre-processed and subsequently the features such as Count as well as the TF-IDF Values are extracted. Like, it was done in the first phase, the extracted features are then utilized for training through the five different machine learning models such as NB, LR, RF, SVM and SGD. Among the five models, the best model is chosen based upon its efficiency as the final classification model to determine the trust worthiness of the web page content. Once the trust worthiness of the website is examined through the machine learning algorithms based on its respective features, the respective website can be termed as phishing or not.

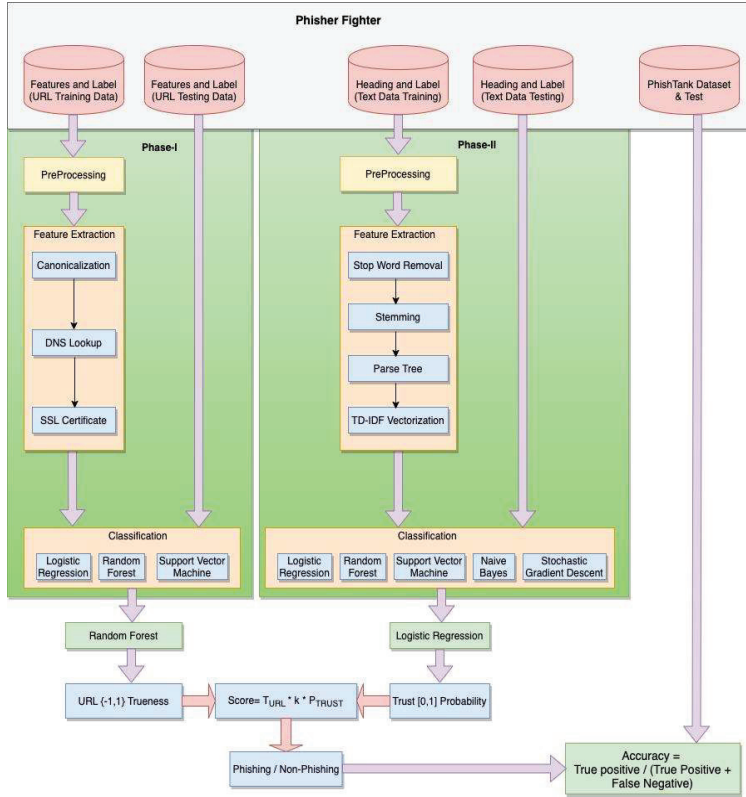


Figure 1 Overall architecture of phisher fighter.

4 Phase-1 URL Analysis

4.1 Pre Processing

The pre-processing process consists of removing the unnecessary columns in the training dataset. It also involves filling up of the lost values in the respective dataset. Once the unwanted columns are eliminated as well as the missing values are filled up, those respective dataset is utilised for extracting the required features. As part of feature extraction, the features from canonicalization, DNS lookup and SSL certificate is extracted.

4.2 Feature Extraction

Various features of the given URL are taken into consideration, the features primarily can be classified into four different classes such as Address Bar,

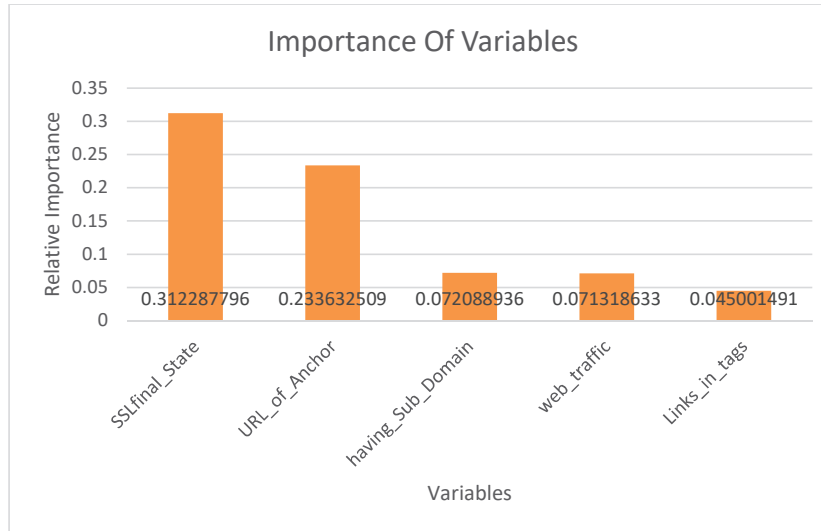


Figure 2 Sample URL features.

Abnormality, HTML and JavaScript as well as Domain based characteristics. The dataset provides us with various DNS based features for prediction of the label. The importance of various features in our dataset is evaluated and plotted in Figure 2.

Canonicalization is a process through which the best URL is picked up eliminating the duplicates. Usually the picked URLs will redirect the user to the respective website homepage. This process plays a major role in optimizing the search engine. From an overall perspective, a DNS lookup is an interaction through which a DNS record is retrieved from a DNS server. The devices that are interconnected realize how to realize the email and domain names of individuals usage into mathematical addresses. An SSL certificate is a significant aspect in the verification process by a browser. Basically, the browsers always verify the certificate credentials such as the issuer signature. Essentially, the browsers verifies the certificate's integrity, validity as well as the revocation status.

4.3 Classifiers

In the classification module, the training dataset is employed to train the classifiers. Three ML classification methods such as LR, RF and SVM are utilized to train up with the extracted URL features. After the training, the

classifier model that fits the best for the respective dataset is selected for further processing. The best fit is measured based upon the classification accuracy of each classifier. In the first phase, with the obtained dataset, Random Forest classifier provides the best classification accuracy of 96% and thereby this particular classifier is designated for further processing.

5 Phase-2 Content Analysis

5.1 Pre-processing

This section comprises of all the pre-processing functionalities necessary to work upon with all the input documents as well as texts. Initially, all required data files are read for training, testing and validation and subsequently pre-processing such as tokenizing, stop word removal, stemming and TF-IDF vectorization. Also, the response variable distribution and data quality audit such as null or lost values are performed. In essence, stop words are employed words like “the”, “a”, “an” and “in” where a search engine uses those words for query searching. These stop words leads to an unnecessary memory consumption and henceforth such words are removed before processing. Stemming is a significant phase of natural language processing pipeline. Basically, it is an approach of generating the morphological variants of the base word. The input to the stemmer will be a tokenized word once the stop words are removed.

5.2 TF-IDF Feature and Count Feature

The required features are extracted using the sci-kit learn python libraries. Bag of words, n-grams as well as TF-IDF weighting are utilized to select the features from the extracted features. Also, word2vec and POS tagging are utilized to obtain the characteristics.

The process of converting an arbitrary text into a fixed length vectors by calculating the number of times each word appears is called as bag-of-words. Basically, N-grams is a series of co-occurring words within a given window. This N-grams are widely utilized in text mining and natural language processing. The TF-IDF weight is basically a statistical measure that is employed to weigh how far a word is influential to its document in a corpus. The weightage augments proportionally to the number of times a word shows up in the particular document. TF-IDF can also can be utilized for stop-word filtering. Word2Vec is basically a two-layer neural network that vectorizes words. It takes text corpus as an input and provides a set of vectors

as output. This processed set of vectors are considered as the feature vectors that represents the words in that particular corpus. Word2Vec can also convert the text in to a numerical format, that can be utilized with the deep neural networks. Part-of-Speech (POS) Tagging is employed to catalogue the words in a text with in the respective corpus based on a specific part of speech. The POS tags characterize the structure is lexical terms within a text or sentence.

5.3 Classifiers

For predicting the phishing, few classifiers are built. The obtained features are sent to various classifiers for further processing. Here classifiers such as NB, LR, Linear SVM, SGD and RF classifiers are utilized. F1 score of each model is compared with each other as soon as fitting with the respective model and the best model is selected as the candidate model and will be selected for the phishing classification. Parameter tuning is performed through GridSearchCV technique on to the candidate model. Moreover, subsequently the best achieving criterion is chosen for the respective classifier. Essentially, GridSearchCV is utilized to fine tune the utilized parameters in the selected model. Among all the classifiers executed in this phase-2, it is noted that the logistic regression provides the maximum efficiency of 75% and eventually designated for further processing.

The best and effective classification models from each phase with highest accuracy is selected for further processing. The phase-1 is the URL analysis and the phase-2 is the content analysis. In phase-1, random forest is designated as the best and effective classifier with 96% of efficiency. In phase-2, logistic regression is designates as the effective and best classifier with 76% of efficiency.

6 Phase-3: Phishing Detection

In this third phase, the detection of website phishing is calculated based upon a score by combining values obtained from the URL analysis as well as the content analysis. Here, both the values obtained from the URL analysis and the content analysis are combined with the AND operation. But, based upon the predictions made and based upon its accuracy, the AND operator can be replaced by a constant value, k that ranges between 0 and 1. The score is calculated based on the given formula,

$$\text{Score} = T_{\text{URL}} * k * P_{\text{TRUST}} \quad (1)$$

where, T_{URL} is the trueness of the URL analysis, P_{TRUST} is the probability of trust of the content analysis and k is the constant. Due to the lack of dataset to predict the value of k , a universal constant of value 1 is assigned to the constant k . The URL trueness varies from -1 to 1 and the trust probability varies between 0 and 1 .

7 Datasets Used

7.1 URL Analysis

An UCI dataset of 11055 data is utilized from the UCI ML [32] repository. The dataset comprises of 32 columns, along with 30 features as well as 1 target as shown in Table 1. In total, the respective repository contains 2456 comments or observations. In order to fit-in the models on to the dataset, the corresponding datasets are partitioned into training sets and testing sets. The partitioned ratio is considered to be 75–25, where 75% is of training dataset and the rest is of testing dataset.

7.2 Content Analysis

The source of the dataset employed for content analysis is the LIAR [33] dataset which is a benchmarked dataset for fake news detection. It comprises of 3 files in .tsv file format for testing, training and validation. The primary dataset contain 13 variables per columns for training, testing and verification. In order to simplify the process, merely 2 variables from the primary dataset is chosen for cataloguing. The remaining variables can be added in due course to increase some more intricacy so that the features can be improvised subsequently. The developed system utilizes three datasets in two columns. The first column contains the New headlines/Text as statements and the second column contains Label class either True/False as Label. The newly synthesized dataset has merely two classes when compared to the six classes from the primary classes. The dataset convention used for reducing the number of classes is shown in Table 3. Also, the dataset employed for this research work is in .csv file format named train.csv and test.csv.

8 Result Analysis

The algorithm when tested with a sample of 4800 real world URLs using the data from PhishTank [30] website using the above mentioned algorithm we

Table 1 UCI phishing dataset [32]

UCI Phishing Dataset		
Attributes	Data Type	Data Range
ID	Integer	$[1, \infty)$
Having IP Address	Integer	$\{-1, 1\}$
URL Length	Integer	$\{-1, 1\}$
Shortening Service	Integer	$\{-1, 1\}$
Having At Symbol	Integer	$\{-1, 1\}$
Double slash redirecting	Integer	$\{-1, 1\}$
Prefix Suffix	Integer	$\{-1, 1\}$
Having Sub Domain	Integer	$\{-1, 0, 1\}$
SSL final State	Integer	$\{-1, 1\}$
Domain registration length	Integer	$\{-1, 1\}$
Favicon	Integer	$\{-1, 1\}$
Port	Integer	$\{-1, 1\}$
HTTPS token	Integer	$\{-1, 1\}$
Request URL	Integer	$\{-1, 1\}$
URL of Anchor	Integer	$\{-1, 0, 1\}$
Links in tags	Integer	$\{-1, 0, 1\}$
SFH	Integer	$\{-1, 1\}$
Submitting to email	Integer	$\{-1, 1\}$
Abnormal URL	Integer	$\{0, 1\}$
Redirect	Integer	$\{-1, 1\}$
On mouseover	Integer	$\{-1, 1\}$
Right Click	Integer	$\{-1, 1\}$
Popup Window	Integer	$\{-1, 1\}$
Iframe	Integer	$\{-1, 1\}$
Age of domain	Integer	$\{-1, 1\}$
DNS Record	Integer	$\{-1, 1\}$
Web traffic	Integer	$\{-1, 0, 1\}$
Page Rank	Integer	$\{-1, 1\}$
Google Index	Integer	$\{-1, 1\}$
Links pointing to page	Integer	$\{-1, 0, 1\}$
Statistical report	Integer	$\{-1, 1\}$
Result	Integer	$\{-1, 1\}$


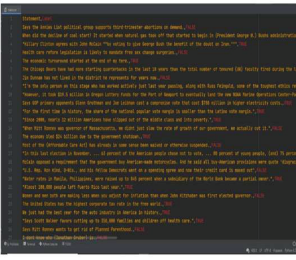
were able to achieve a True Positive of 90.68%. The average computation time for the sample was evaluated to be 10.4 seconds.

The URL analysis efficiency chart is shown in Figure 3. Various efficiency metrics are employed over three different ML techniques such as LR, RF and

Table 2 LIAR dataset [33]

Attributes	Data Type
ID	Integer
Label	String
Statement	String
Subject(s)	String
Speaker	String
Speaker's job title	String
State info	String
Party affiliation	Integer
Barely true counts	Integer
False counts	Integer
Half true counts	Integer
Mostly true counts	Integer
Pants on fire counts	Integer
Location of the speech	String

Table 3 Dataset convention and the screenshots of dataset used

Original	New	Raw Dataset (test.csv)	Structured Data (test.csv)
True	True		
Mostly-True	True		
Half-True	True		
Barely-True	False		
False	False		
Pants-fire	False		

SVM. After the respective analysis, it is observed and noted that RF provides the maximum accuracy with 97% and LR provided an accuracy of 92%.

The content analysis efficiency chart is depicted in Figure 4. Efficiency metrics are worked out on five different ML methods such as NB, LR, SVM, SGD and RF. After the metric analysis, it is noticed that LR provides the maximum accuracy with 76% and SGD provided an accuracy of 56%.

A comparison of the proposed system along with various existing detection technique is shown in Table 4. Initially we observed that, Large-scale automatic classification of pages [21], EBDIS [22], PhishCatch [23], CANTINA [24] and Web Phishing Detection using a deep learning framework [31] performs poorly as they have higher false negative rate. Despite the

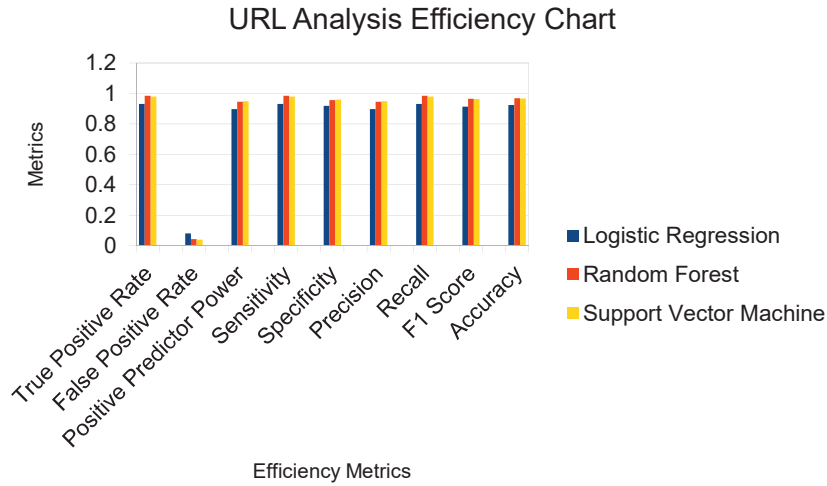


Figure 3 Efficiency Chart of the URL analysis.

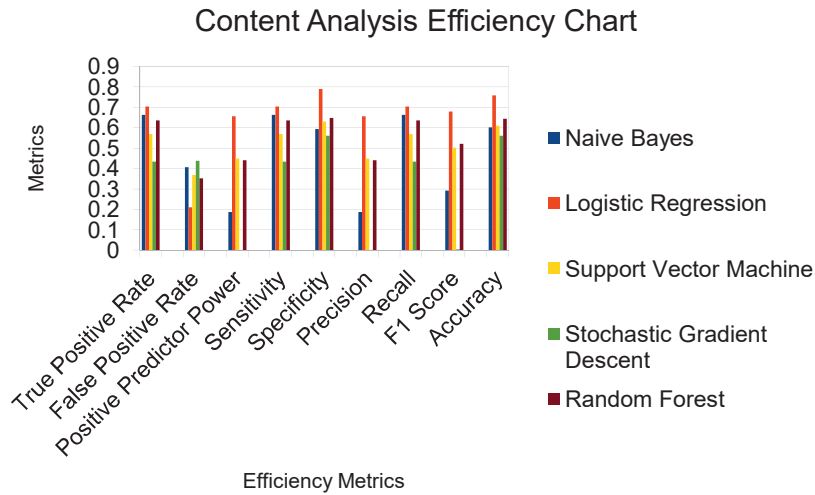


Figure 4 Efficiency Chart of the Content analysis.

fact that the novel phishing approach using auto updated whitelist [25] and Web Phishing Detection based on graph mining [26] have less false negative rate, their running time is significantly higher than that of the proposed methodology. The proposed system was evaluated with a dataset containing 520 elements which gave a True Positive percentage of 97.6% with a 3.65% False Negative rate. This respective execution took an overall computation

Table 4 Phisher fighter vs existing techniques

Detection Technique	False Negative (FN) (%)
Large-scale automatic classification of pages [21]	16–30
EBDIS [22]	25
PhishCatch [23]	20
CANTINA [24]	11
Novel approach using auto updated whitelist [25]	1.48
Web Phishing Detection Based on Graph Mining [26]	3
Web Phishing Detection Using a Deep Learning Framework [31]	10.8
Phisher Fighter (Proposed Method)	9.31

time of 4 hours. Out of 4800 tuples, 4335 of the phishing websites have been classified correctly and 447 have been diagnosed wrong. The average time taken for single verification was approximated to be 12.1126 seconds. When the dataset was scaled up to 4800 elements, the efficiency came down with a True Positive percentage of 90.68% and a False Negative of 9.31%. The time yielded to compute is about 10 hrs.

9 Conclusion and Future Work

An effective method to detect phishing in websites based upon on the URL analysis and content analysis with TF IDF values of the content of the respective website is proposed and implemented. On executing the proposed method Phisher Fighter, an efficiency of 90.68% is achieved with an execution of about 10 hours. The time complexity is equated to $\text{LOG}(n,2)$. The phisher fighter utilized 4800 elements for its execution. Another existing phishing detecting system called the Cantina is able to perform with an efficiency of 76.20% and it is observed that the proposed system, Phisher Fighter is far superior than Cantina in terms of efficiency.

The efficiency of this developed system can be improvised once the dataset for training is more. Also, feature selection method like topic modelling can be introduced and implemented for better and accurate results. The constant value of k can be estimated by an appropriate mathematical model or by an empirical method for better handling of data and effective outcome. The bottleneck in computing time could be improved by parallelizing the algorithm in recursive decomposition.

This work can be replicated by exploring more on appropriate larger datasets and subsequently employ deep learning methods and artificial intelligence for better efficacy in detection.

References

- [1] Zhuang, W., Jiang, Q., and Xiong, T. (2012, June). An intelligent anti-phishing strategy model for phishing website detection. In *2012 32nd International Conference on Distributed Computing Systems Workshops* (pp. 51–56). IEEE.
- [2] Alkhozai, M. G., and Batarfi, O. A. (2011). Phishing websites detection based on phishing characteristics in the webpage source code. *International Journal of Information and Communication Technology Research*, 1(6).
- [3] Sahingoz, O. K., Buber, E., Demir, O., and Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, 117, 345–357.
- [4] Ali, W. (2017). Phishing website detection based on supervised machine learning with wrapper features selection. *International Journal of Advanced Computer Science and Applications*, 8(9), 72–78.
- [5] Sankhyan, R., Shetty, A., Dhanopia, L., Kaspale, C., and Dantal, P. G. (2018). PDS-Phishing Detection Systems. *Safety*, 5(04).
- [6] Rao, R. S., and Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications*, 31(8), 3851–3873.
- [7] Varshney, G., Misra, M., and Atrey, P. K. (2016). A survey and classification of web phishing detection schemes. *Security and Communication Networks*, 9(18), 6266–6284.
- [8] Hara, M., Yamada, A., and Miyake, Y. (2009, March). Visual similarity-based phishing detection without victim site information. In *2009 IEEE Symposium on Computational Intelligence in Cyber Security* (pp. 30–36). IEEE.
- [9] Bergholz, A., Paaß, G., D’Addona, L., and Dato, D. (2010). A real-life study in phishing detection. In *Proceedings of the conference on email and anti-spam (CEAS)* (Vol. 1, pp. 1–10).
- [10] Subasi, A., Molah, E., Almkallawi, F., and Chaudhery, T. J. (2017, November). Intelligent phishing website detection using random forest classifier. In *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)* (pp. 1–5). IEEE.
- [11] Afroz, S., and Greenstadt, R. (2009, September). Phishzoo: An automated web phishing detection approach based on profiling and fuzzy matching. In *Proc. 5th IEEE Int. Conf. Semantic Comput.(ICSC)* (pp. 1–11).

- [12] Ubung, A. A., Jasmi, S. K. B., Abdullah, A., Jhanjhi, N. Z., and Supramaniam, M. (2019). Phishing Website detection: An improved accuracy through feature selection and ensemble learning. *International Journal of Advanced Computer Science And Applications*, 10(1), 252–257.
- [13] Ali, W. (2017). Phishing website detection based on supervised machine learning with wrapper features selection. *International Journal of Advanced Computer Science and Applications*, 8(9), 72–78.
- [14] Basnet, R. B., Sung, A. H., and Liu, Q. (2012, June). Feature selection for improved phishing detection. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 252–261). Springer, Berlin, Heidelberg.
- [15] Alnajim, A., and Munro, M. (2009, April). An anti-phishing approach that uses training intervention for phishing websites detection. In *2009 Sixth International Conference on Information Technology: New Generations* (pp. 405–410). IEEE.
- [16] Jain, A. K., and Gupta, B. B. (2019). A machine learning based approach for phishing detection using hyperlinks information. *Journal of Ambient Intelligence and Humanized Computing*, 10(5), 2015–2028.
- [17] He, M., Horng, S. J., Fan, P., Khan, M. K., Run, R. S., Lai, J. L., . . . and Sutanto, A. (2011). An efficient phishing webpage detector. *Expert systems with applications*, 38(10), 12018–12027.
- [18] Li, Y., Yang, Z., Chen, X., Yuan, H., and Liu, W. (2019). A stacking model using URL and HTML features for phishing webpage detection. *Future Generation Computer Systems*, 94, 27–39.
- [19] Chiew, K. L., Choo, J. S. F., Sze, S. N., and Yong, K. S. (2018). Leverage website favicon to detect phishing websites. *Security and Communication Networks*, 2018.
- [20] Dadgar, S. M. H., Araghi, M. S., and Farahani, M. M. (2016, March). A novel text mining approach based on TF-IDF and Support Vector Machine for news classification. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)* (pp. 112–116). IEEE.
- [21] Whittaker, C., Ryner, B., and Nazif, M. (2010). Large-scale automatic classification of phishing pages.
- [22] Stone, A. (2007). Natural-language processing for intrusion detection. *Computer*, 40(12), 103–105.
- [23] Yu, W. D., Nargundkar, S., and Tiruthani, N. (2009, July). Phishcatch – a phishing detection tool. In *Proceedings of the 2009 33rd Annual IEEE International Computer Software and Applications Conference – Volume 02* (pp. 451–456).

- [24] Zhang, Y., Hong, J. I., and Cranor, L. F. (2007, May). Cantina: a content-based approach to detecting phishing web sites. In *Proceedings of the 16th international conference on World Wide Web* (pp. 639–648).
- [25] Jain, A. K., and Gupta, B. B. (2016). A novel approach to protect against phishing attacks at client side using auto-updated white-list. *EURASIP Journal on Information Security*, 2016(1), 1–11.
- [26] Futai, Z., Yuxiang, G., Bei, P., Li, P., and Linsen, L. (2016, October). Web phishing detection based on graph mining. In *2016 2nd IEEE international conference on computer and communications (ICCC)* (pp. 1061–1066). IEEE.
- [27] HR, M. G., Adithya, M. V., and Vinay, S. (2020). Development of anti-phishing browser based on random forest and rule of extraction framework. *Cybersecurity*, 3(1), 1–14.
- [28] Alkhalil, Z., Hewage, C., Nawaf, L., and Khan, I. (2021). Phishing Attacks: Recent Comprehensive Study and a New Anatomy. *Frontiers in Computer Science*, 3, 6.
- [29] Vrbančič, G., Fister Jr, I., and Podgorelec, V. (2020). Datasets for phishing websites detection. *Data in Brief*, 33, 106438.
- [30] <http://data.phishtank.com/data/online-valid.csv>
- [31] Yi, P., Guan, Y., Zou, F., Yao, Y., Wang, W., and Zhu, T. (2018). Web phishing detection using a deep learning framework. *Wireless Communications and Mobile Computing*, 2018.
- [32] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.
- [33] Wang, W. Y. (2017). “liar, liar pants on fire”: A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648.

Biographies



E. Sri Vishva is currently pursuing the bachelor's degree in Computer science and Engineering Vellore Institute of Technology, Vellore. He is currently a junior student in the School of Computer Science and Engineering. His research areas include Information Security and Machine Learning.



D. Aju received his PhD. in Computer Science and Engineering from Vellore Institute of Technology, Vellore, India. He received his M.Tech. degree in Computer Science and IT from Manonmaniam Sundaranar University, Tirunelveli. He received his M.C.A degree from Madras University, India. Presently, he is working as Associate Professor at Vellore Institute of Technology in the department of Information Security, School of Computer Science and Engineering. He has published more than 30 research articles in different reputed international peer-reviewed journals. And, he has served as reviewer for few international peer-reviewed journals. He is having more than 16 years of teaching and research experience. Consecutively, he has received research awards from 2014 to 2019 for his outstanding contribution towards research and publication at Vellore Institute of Technology. His research area of interest includes Digital Image Processing, Medical Imaging, Computer Graphics, Cyber Security and Digital Forensics.