ISSN [Online]: 2583-2654

Detecting Phishing Website Using Machine Learning

Mercy Flora A[1], Cladina Sharon R[2], Nicky M[3], Rebecca Ann Dingle[4], Rithika K[5]
Department of Computer Science & Engineering,
Dr.T.Thimmaiah Institute of Technology,
India, VTU

Abstract: Trying to gather personal information through deceptive ways is becoming more common nowadays. In order to assist the user to be aware of the access to such websites, the implemented system notifies the user through email and also pop-up, when trying to access a phishing site. Many blacklisted websites has been published to appear as an original site in order to trap user by asking them to input their personal details. For example password, bank account, email address etc. Machine Learning was implemented to develop this proposed system. A Machine learning technique identifies phishing URLs typically assess a URL based on some feature or set of features extracted from it. A pop up notification will be displayed when user clicks on the blacklisted URL. A message from admin will be displayed once user clicks "OK" from the pop – up notification and also a message from admin will be displayed as user received the Gmail notification, so that individuals can be alerted while browsing or accessing a particular website. Therefore, it can be utilized for identification and authentication and become a legitimate tool to prevent an individual from getting tricked.

Keywords:Blacklisted, phishing, alert, pop-up notification, Email notification, Feature Extraction, Machine Learning.

I. INTRODUCTION

Phishing can be defined as impersonating a valid site to trick users by stealing their personal data comprising usernames, passwords, accounts numbers, national insurance numbers, etc. Phishing frauds might be the most widespread cybercrime used today. There are countless domains where phishing attack can occur like online payment sector, webmail, and financial institution, file hosting or cloud storage and many others. The webmail and online payment sector was embattled by phishing more than in any other industry sector. Phishing can be done through email phishing scams and spear phishing hence user should be aware of the consequences and should not give their 100 percent trust on common security application. Machine Learning is one of the efficient techniques to detect phishing as it removes drawback of existing approach.

The objectives which is the most vital thing in proposed project is to verity the validity of the website by capturing blacklisted URLs. To notify the user on blacklisted website through pop-up while they are trying to access and to notify the user on blacklisted website through email while they are trying to access. This proposed project will allow administrator to add blacklisted URL's in order to alert user during their inquiry.

The two scope of project, which is well known as user scope and system scope. User has some responsibility towards the system. The system includes a few standards and policies that requires to be obliged in order to comply the system. The user can be notified if blacklisted website is being accessed. The admin can capture the blacklisted URL's to alert user. The system involves features like capturing blacklisted website, viewing blacklisted website, displaying pop-up notification and also displaying email notification.

ISSN [Online]: 2583-2654

II. DATASET

URLs of benign websites were collected from www.alexa.com and The URLs of phishing websites were collected from www.phishtank.com. The data set consists many URLs which include benign URLs and phishing URLs. Benign URLs are labelled as "0" and phishing URLs are labelled as "1".

III.FEATURE EXTRACTION

We have implemented python program to extract features from URL. Below are the features that we have extracted for detection of phishing URLs.

1) **Presence of IP address in URL:** If IP address present in URL then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.

2) **Presence of @ symbol in URL**: If @ symbol present in URL then the feature is set to 1 else set to 0. Phishers add special symbol @ in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol [4].

3) **Number of dots in Hostname**: Phishing URLs have many dots in URL. For example http://shop.fun.amazon.phishing.com, in this URL phishing.com is an actual domain name, whereas use of "amazon" word is to trick users to click on it. Average number of dots in benign URLs is 3. If the number of dots in URLs is more than 3 then the feature is set to 1 else to 0.

4) **Prefix or Suffix separated by (-) to domain:** If domain name separated by dash (-) symbol then feature is set to 1 else to 0. The dash symbol is rarely used in legitimate URLs. Phishers add dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example Actual site is http://www.onlineamazon.com but phisher can create another fake website like http://www.online-amazon.com to confuse the innocent users.

5) **URL redirection:** If "//" present in URL path then feature is set to 1 else to 0. The existence of "//" within the URL path means that the user will be redirected to another website.

6) **HTTPS token in URL:** If HTTPS token present in
URL then the feature is set to 1 else to 0. Phishers may add the "HTTPS" token to the domain part of a URL in order to trick users. For example, http://https-www-paypal-it-mpp-home.soft-hair.com.

7) **Information submission to Email:** Phisher might use "mail()" or "mailto:" functions to redirect the user's information to his personal email[4]. If such functions are present in the URL then feature is set to 1 else to 0.

8) **URL Shortening Services "TinyURL":** TinyURL service allows phisher to hide long phishing URL by making it short. The goal is to redirect user to phishing websites. If the URL is crafted using shortening services (like bit.ly) then feature is set to 1 else 0

9) **Length of Host name:** Average length of the benign URLs is found to be a 25, If URL's length is greater than 25 then the feature is set to 1 else to 0

10) **Presence of sensitive words in URL:** Phishing sites use sensitive words in its URL so that users feel that they are dealing with a legitimate webpage. Below are the words that found in many phishing URLs :- 'confirm', 'account', 'banking', 'secure', 'ebyisapi', 'webscr', 'signin', 'mail', 'install', 'toolbar', 'backup', 'paypal', 'password', 'username', etc;

11) **Number of slash in URL:** The number of slashes in benign URLs is found to be a 5; if number of slashes in URL is greater than 5 then the feature is set to 1 else to 0.

12) **Presence of Unicode in URL:** Phishers can make a use of Unicode characters in URL to trick users to click on it. For example the domain

ISSN [Online]: 2583-2654

“xn--80ak6aa92e.com” is equivalent to "apple.com". Visible URL to user is "apple.com" but after clicking on this URL, user will visit to “xn--80ak6aa92e.com” which is a phishing site.

13) **Age of SSL Certificate:** The existence of HTTPS is very important in giving the impression of website legitimacy [4]. But minimum age of the SSL certificate of benign website is between 1 year to 2 year.

14) **URL of Anchor:** We have extracted this feature by crawling the source code oh the URL. URL of the anchor is defined by <a> tag. If the <a> tag has a maximum number of hyperlinks which are from the other domain then the feature is set to 1 else to 0.

15) **IFRAME:** We have extracted this feature by crawling the source code of the URL. This tag is used to add another web page into existing main webpage. Phishers can make use of the “iframe” tag and make it invisible i.e. without frame borders [4]. Since border of inserted webpage is invisible, user seems that the inserted web page is also the part of the main web page and can enter sensitive information.

16) **Website Rank:** We extracted the rank of websites and compare it with the first One hundred thousand websites of Alexa database. If rank of the website is greater than 10,0000 then feature is set to 1 else to 0.

IV.RELATED WORK

In emerging technology industry which deeply influence today’s security problems has given a non-ease of mind to some employer and home users. Occurrences that exploit human vulnerabilities have been on the upsurge in recent years. [1] In the dimension of new era there are many security systems being developed to ensure security is given the utmost priority and prevention to be taken from being hacked by those who are involved in cyber-criminal and essential prevention is also taken as high consideration in organization to ensure network security is not being breached. Cyber security employees are currently searching for trustworthy and steady detection techniques for phishing websites detection. [2] Due to wide usage of internet to perform various activities such as online bill payment, banking transaction, online shopping, and, etc. Customers face numerous security threats like cybercrime. There are many cybercrime that are extensively executed for example spam, fraud, cyber terrorisms and phishing. Among this phishing is known as the popular cybercrime today. [3] Phishing has become one amongst the highest 3 most current forms of law-breaking in line with recent reports, and both frequency of events and user susceptibleness has enlarged in recent years, more combination the danger of economic damage. [4]

Phishing is a type of practice done on the Internet where individual data are obtained by illegal approaches.[5] It supply of obtaining sensitive information, as an example, usernames, passwords, and positive identification points of interest, often for malignant reasons, by taking up the looks of an electronic correspondence. Phishing attack will be enforced in varied kind like Email phishing, web site phishing, spear phishing, Whaling, Tab off his guard, Evil twin phishing etc. [6] Phishing is known as webpage violence. [7] Phishing is often done by email spoofing or texting, and it typically guides user to enter points of interest at a fake web site which look and feel the same. It tries to handle the increasing range of phishing got to be met by clients in awareness and alternative efforts to ascertain protection numerous anti-phishing tools. A number of sites have currently created optional instruments for applications, like maps for redirection but clients ought to not utilize similar passwords anywhere on the net. [8] The primary key feature is to allow user to inquire whether visited websites is original or fake.

This paper proposes a security tool called as Detecting Phishing Website Using Machine Learning to detect and eliminate phishing sites.

ISSN [Online]: 2583-2654

V. LITERATURE REVIEW

The current situation that is majority of the population has been fooled into giving their

personal details to hackers without noticing it. Many blacklisted website has been publish to appear as an original site in order to trap user by asking them to input their personal details. For example, password, bank account, email address and etc. Phishing activity in early 2016 was the highest ever recorded since it began monitoring in 2004.The total number of phishing attacks in 2016 was 1,220,523. Tis was a 65 percent increase over 2015. In the fourth quarter of 2004, there were 1,609 phishing attacks per month. In the fourth quarter of 2016, there was an average of 92,564 phishing attacks per month, an increase of 5,753% over twelve years. [9] According to the Anti-Phishing Working Group (APWG), there are at least 47, 324 phishing attacks and a top-ten American bank estimates that at least US$300 is lost for every hour that a phishing site remains up. [10] Machine learning is that the science of obtaining computers to act while not being expressly programmed. [11] Machine Learning was implement to develop this proposed system. Machine learning techniques identifies phishing URLs typically assess a URL based on some feature or set of features extracted from it. [12].Thus, before coming to conclusion that this was the major problem, related products were examined and compared view their libation before progressing to the proposed project.

Phishtank was proposed to carry out the inspection once a link has been pasted on the section given. This allow user to keep on track of faked website. They can copy and paste the link in order to identify whether the site that they are going to access is safe or not safe. User can use the website search feature directly or they can use information from PhishTank through its API. A search engine displayed on PhishTank website is to be used as the first method. Using its API will be the second method. API service can be avail by software builder after registering themselves on PhishTank website. Both methods mentioned above do not cost a single penny. The purpose of API's usage is for user who has basis information on software development.

Limitation of this project is there was no facility of displaying pop-up and email notification once user had access blacklisted website. [13]

PhishZoo was proposed to evaluate a new method for web phishing detection based on profiles of complex sites' appearance and content. PhishZoo makes profiles of sites comprising of the website contents and images displayed. These profiles are kept in a local folder and are either synchronized against the newly loaded sites at the time of loading or against risky sites for instance, links in email offline. Limitation of this project is there was no facility of displaying pop-up and email notification once user had access blacklisted website. [14].

GoldPhish was proposed to perceive and report phishing sites. This was done by using optical character recognition (OCR) to recite the text from an image of the page precisely. From the company logo, grasping the top hierarchical areas from a search engine, and comparing them with the current web site. The forte of the tool lies in the user's capability to recognize famous company logos. A phishing site cannot change a familiar company logo without the phishing target perceiving. Limitation of this project is there was no facility of displaying pop-up and email notification once user had access blacklisted website. [15]

VI. MACHINE LEARNING ALGORITHEMS

Two machine learning classification model Random forest and Naïve bayes has been selected to detect phishing websites.

6.1 Random Forest Algorithm

Random forest algorithm is one of the most powerful algorithms in machine learning

ISSN [Online]: 2583-2654

technology and it is based on concept of decision tree algorithm. Random forest algorithm creates the forest with number of decision trees. High number of tree gives high detection accuracy.

Creation of trees are based on bootstrap method. In bootstrap method features and samples of dataset are randomly selected with replacement to construct single tree. Among randomly selected features, random forest algorithm will choose best splitter for the classification and like decision tree algorithm; Random forest algorithm also uses gini index and information gain methods to find the best splitter. This process will get continue until random forest creates n number of trees.

Each tree in forest predicts the target value and then algorithm will calculate the votes for each predicted target. Finally random forest algorithm considers high voted predicted target as a final prediction.

6.2 Naïve Bayes Classifier Algorithm

Naive Bayes is a simple technique for constructing classifiers models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

The problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting.

Web pages containing more external links than internal ones and password field input are classified as suspicious. A website content with more external links than internal links is an attempt to achieve some similarities and styles from external sources with the objective to steal user credential.

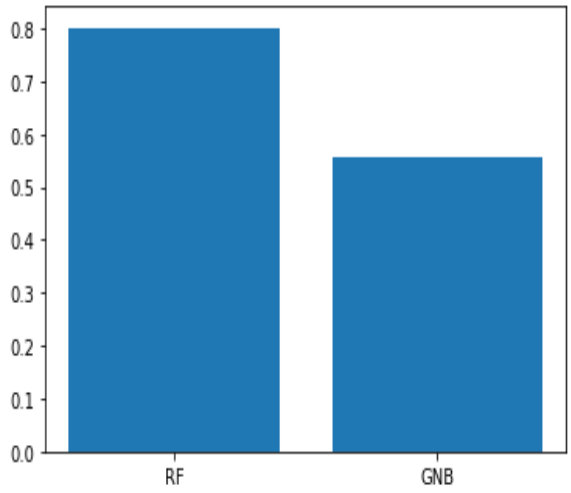

Fig. Detection accuracy comparison

VII.PROPOSED METHOD

7.1 System Requirements

Hardware requirements

System : Intel i3 2.1 GHZ
Memory : 4GB.
Hard Disk : 80 GB.
Android Phone

Software requirements

Operating System : Windows 7 / 8/10.
Language : python(3.7.4)

Phishing is a type of practice done on the Internet where individual data are obtained by illegal approaches. It supply of obtaining sensitive information, as an example, usernames, passwords, and positive identification points of interest, often for malignant reasons, by taking up the looks of an electronic correspondence.

This project was developed in Python. Python has a huge set of libraries and extensions, which can be easily used in Machine Learning. Python librariesare the best source for machine learning algorithms where nearly all types of machine learning algorithms are readily available for

ISSN [Online]: 2583-2654

Python, thus easy and quick evaluation of ML algorithms is possible.

VIII. Implementation And Result

Scikit-learn tool has been used to import Machine learning algorithms. Dataset is divided into training set and testing set in 80:20 ratio respectively. Each classifier is trained using training set and testing set is used to evaluate performance of classifiers. Performance of classifiers has been evaluated by calculating classifier's accuracy score, false negative rate and false positive rate.

Result shows that Random forest algorith1m gives better detection accuracy which is 80.14 with lowest false negative rate than Naïve Bayes algorithm.

.

Result also shows that detection accuracy of phishing websites increases as more dataset used as training dataset. All classifiers perform well when 80% of data used as training dataset.

Fig. above show the detection accuracy of both the classifiers when 80% of data used as training dataset and graph clearly shows that detection accuracy increases and random forest detection accuracy is maximum than other classifier.

8.1 Description Of Output

Based on **Fig 1,** this is the main page, where user can copy paste an URL in the space provided in order to know whether the given URL is legitimate or phishing.

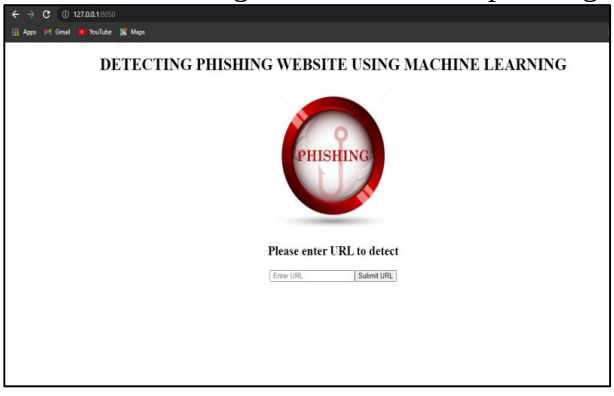

Fig 1: Admin Page

Based on **Fig 2,** Once the URL is submitted, it shows whether the given URL is legitimate or phishing, if the given URL is a original link then it will redirect to the actual website.

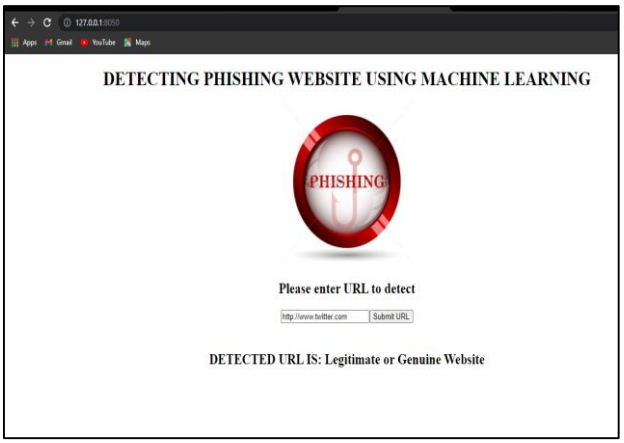

Fig 2: Legitimate Site Detected

Fig 3 shows that a pop up notification a when user clicks on the blacklisted URL. The pop-up notification is an alert box to apprise the user by questioning whether they wish to continue knowing that it may be a phishing site.

Fig 3: POP-UP Notification

Based on **Fig 4**, user receives a Gmail notification once user clicks “YES” from the pop–up. This mail will notify the user that the site accessed is confirmed a phishing site.

ISSN [Online]: 2583-2654

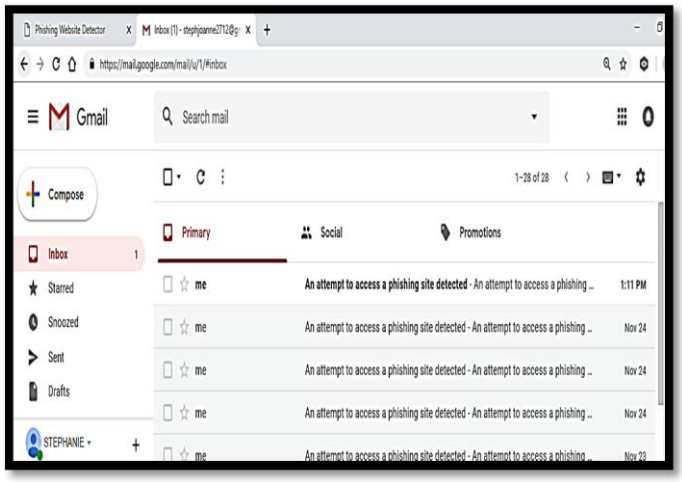

Fig 4: Gmail Notification

Fig 5 shows a screenshot of Gmail interface. A notification from Gmail will be apprised once user clicks “YES" from the pop-up notification.

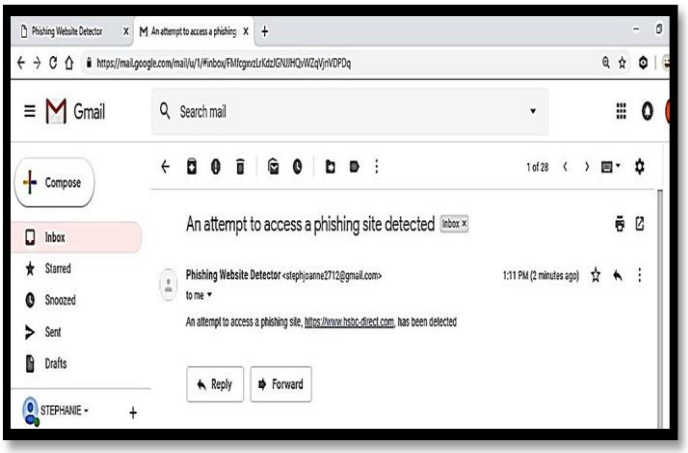

Fig 5: Message From Admin Through Email

As the overall discussion, the proposed project has been successfully coded and developed, and has met the expectations made during the project proposal phase. From the various results aggregated during the various tests, the system can be concluded that it has been successfully developed to its specifications. The system interface has been successfully designed, the functions are coded into function and the different requirements are met.

IX.CONCLUSION

After reviewing and researching for appropriate monitoring tools, proposed system has been identified and chosen to address the complexity of monitoring requirement for current situation. This software is designed to show awareness of the extensive level of its functionality, features that can be displayed in the monitoring era. The system fosters many features in comparison of other software. Its unique features such as capturing blacklisted URL’s from the browser directly to verify the validity of the website, notifying user on blacklisted websites while they are trying to access through pop-up, and also notifying through email. This system will assist user to be alert when they are trying to access a blacklisted website.

In conclusion, this system is designed for resources are used as intended, prevents from valuable information from leaks out, produce better control mechanism and alerts the user to keep their private information safe. Like any other programs, there are improvements which could be made into this system. Based on the capabilities which the current system processes, text message integration would a great recommendation that could be made to improve the program in the future. The future version of the application could also implement an option to directly notify the blacklisted website with a text message. The program could be made to access the list as an attachment. This text message integration function would further the usabilityof the application.

Acknowledgement

Authors are grateful to Dr.T.Thimmaiah Institute Of Technology and the Faculty of Computer Science and Engineering,, Visvesvaraya Technological University, Karnataka for their support.

REFERENCES

[1] W. Liu, X. Deng, G. Huang, and A. Y. Fu, “An antiphishing strat- egy based on visual similarity assessment,” IEEE Internet Computing, vol. 10, no. 2, pp. 58–65, 2006.

[2] M. Al-diabat, ”Detection and Prediction of Phishing Websites using Classification Mining Techniques”, International Journal of Computer Application, vol. 147, no. 1, pp. 5-11, 2016.

ISSN [Online]: 2583-2654

[3] V. S. Lakshmi, M. S. Vijaya, "Efficient prediction of Phishing Websites Using Supervised Learning Algorithms," Procedia Engineering, vol. 30, no. 4, pp. 798-805, 2012

[4] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and P. S. Yu, "Top 10 algorithm in data mining", Knowledge and Information System , vol. 14, no. 6, pp. 1-37, 2008

[5] W. Hadi, F. Aburrub, and S.Alhawari, "A new fast associative classification algorithm for detecting phishing websites", Applied Soft Computing vol. 48, no. 3, pp. 729-734, 2016.

[6] M.A.U.H Tahir, S. Asghar, A. Zafar, S.Gillani, "A Hybrid Model to Detect Phishing-sites Using Supervised Learning Algorithms," International Conference OnComputational Science and Computational Intelligence(CSCI), vol. 15, no. 10, pp. 1126-1133, IEEE, 2016.

[7] .S. Savag, G. M. Voelker, " Learning to detect malicious URLs", ACM Transactions on Intelligent System and technology, vol. 2, no. 9, pp. 30: 1-30:24, 2011.

[8] S. Purkait, "Phishing counter measures and their effectiveness-literature review", Information Management & Computer Security, vol. 20, no. 5, pp. 382-420, 2012.

[9] N. Abdelhamid, A. Ayesh, F. Thabtah, "Phishing Detection based Associative Classification", Data Mining. Expert System with Application (ESWA), vol. 41, no. 6, pp. 5948-5959,2014.

[10] H. H. Nguyen, D.T., "Nguyen, Machine Learning based Phishing websites detection," Recent Advances in Electrical Engineering and related science, vol 22, no.1, pp.123-131, 2016.

[11] R.S. Rao and S.T.Ali, "PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach," Procedia Computer Science, vol. 54, no. Supplement C, pp. 147-156, 2015.

[12] S. Marchal, J. Franois, R. State, and T. Engel, "PhishStorm: Detecting Phishing With Streaming Analytics," IEEE Transactions on Network and Service Management, vol. 11, no. 4, pp. 358-471, Dec. 2014.

[13] K.-T. Chen, J.-Y. Chen, C.-R. Huang, and C.-S. Chen, "Fighting phish- ing with discriminative keypoint features," IEEE Internet Computing, vol. 13, no. 3, pp. 56–63, 2009.

[14] A. Y. Fu, L. Wenyin, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (emd)," IEEE Trans. Dependable Secur. Comput., vol. 3, no. 4, pp. 301–311, 2006.

[15] R M Mohammad, F.Thabtah, L. McCluskey, :"Tutorial and Critical Analysis of Phishing websites methods," Computer Science Review, vol. 17, no.8, pp. 1-24, 2015.