IJIRIS: International Journal of Innovative Research in Information Security
Volume 09, Issue 03, June 2023
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives

Malevolent WEBSITE identification using Machine Learning Techniques

J Jesy Janet Kumari,
Assistant Professor, Department of CSE,
The Oxford College of Engineering, Bangalore
ebijesy.j@gmail.com

Publication History
Manuscript Reference No: IJIRIS/RS/Vol.09/Issue03/JNIS10101
Research Article | Open Access | Double-blind Peer-reviewed |Article ID: IJIRIS/RS/Vol.09/Issue03/JNIS10101
Received: 02, June 2023 | Accepted: 20, June 2023 | Published: 23, June 2023
Volume 2023 | Article ID JNIS10101 http://www.ijiris.com/volumes/Vol09/iss-03/22.JNIS10101.pdf
ArticleCitation:Jesy(2023).Malevoent Website Identification using Machine Learning Technique. International Journal of Innovative Research in Information Security (IJIRIS), 10, 160-167 doi: https://doi.org/10.26562/ijiris.2023.v0903.22
BibTex key: **Jesy2023Malevolent**

Copyright: ©2023 This is an open access article distributed under the terms of the Creative Commons Attribution License; which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Abstract: Threats to network information security are currently increasing both in number and seriousness. Today's hackers are mostly focused on attacking end-to-end systems and finding weak points in people. These strategies include pharming, phishing, and social engineering. These assaults include a stage where malicious Uniform Resource Locators (URLs) are used to trick visitors. Consequently, harmful URL detection is growing in popularity. The use of machine learning and deep learning algorithms to identify malicious URLs has been proven in several academic articles. In this study, we propose a malicious URL detection system that makes use of machine learning approaches based on proposed URL behaviors and attributes. Bigdata technology is additionally being utilized to enhance the detection of risky URLs based on peculiar behavior. A novel set of URL traits and behaviors, a machine learning algorithm, and big data technologies make up the suggested detection technique. The trials' findings imply that the suggested URL characteristics and behavior can significantly improve the ability to identify potentially harmful URLs. This shows that the suggested method might be regarded as the best and most convenient way to find counterfeit URLs.

I. INTRODUCTION

The Uniform Resource Locator (URL) is a method of referring to Internet resources. Sahooetal. Published information in. the URL's properties and two essential components are protocol identifier, which specifies which protocol should be used, and The IP address or domain name is specified by the resource name. the name of the resource's location It is clear that each The structure and format of a URL are specific Attackers frequently attempt to alter one or more of the URL's structural components to deceive consumers in order to distribute a malicious URL Malicious URLs are links that have a negative impact on users[1].These Users will be sent to resources or pages using URLs. Attackers have the ability to run code on users' computers and divert them. To un welcome web pages, dangerous web pages[2].

According to figures published in, assaults using the distributing malicious URL strategy are ranked #1 among the top ten most popular attack strategies in 2019. According to this report, the three main URL spreading strategies, harmful URLs, bot net URLs, and phishing URLs, are seeing an increase in the number of attacks as well as the level of danger they pose. There are two primary trends in the identification of malicious URLs now malicious URL detection based on signals or sets of rules, and malicious URL detection based on behavior analysis approaches. The method of detecting malicious URLs based on a collection of markers or criteria can detect harmful URLs rapidly and accurately. This approach, on the other hand, is in capable of detecting new dangerous URLs that do not match the specified indicators or guide lines. Machine learning or deep learning algorithms are used to classify URLs based on their actions in the process of detecting malicious URLs based on behavior analysis techniques. Machine learning algorithms are used to classify URLs based on their properties in this article. A new URL attribute is also included in the document information [3]. Machine learning methods are employed in our study to classify URLs based on their properties and behaviors. The traits, which are novel to the literature, are derived from the static and dynamic behavior of URLs. The research's key contribution is the newly proposed features. The entire malicious URL detection system includes machine learning algorithms. Support vector machine (SVM) and Random Fore stare two supervised machine learning methods employed (RF). The following could be a breakdown of the paper's structure. The second section of the paper looks at some recent research on malicious URL detection. In Section III ,the suggested machine learning-based approach for detecting malicious URLs is provided. The new features for the URL detection process are detailed intimately during this section. Section IV contains the experimental results and discussions. Section V brings the paper to a detailed [4].

II. RELATEDWORK

A. Signature based Malicious URL Detection

Long ago, studies on malicious URL detection utilizing signature sets were researched and implemented [5]. Most of these investigations rely on databases of known harmful URLs. A database query is run whenever a new URL is accessed. If a URL is blacklisted, it is assumed to be dangerous, and a warning is issued; otherwise, URLs are assumed to be secure. The biggest drawback of this strategy is that it will be extremely difficult to detect new malicious URLs that are not included in the provided list [6].

B. Machine Learning based Malicious URL Detection.

Supervised learning, unsupervised learning, and semi supervised learning are the three types of machine learning algorithms that can be used to detect dangerous URLs. The methods of detection are based on URL behavior. looked into a variety of malicious URL systems based on machine learning methods. SVM, Logistic Regression, Nave Bayes, Decision Trees, Ensembles, Online Learning, and soon are examples of machine learning algorithms. The two algorithms RF and SVM are employed in this paper. The experimental results will show the accuracy of the set two algorithms with various parameter configurations. URLs can be classified into two classes based on their behaviors and characteristics: static and dynamic [7]. Authors proposed approaches for assessing and extracting static behavior of URLs in their investigations [8] ,including Lexical, Content ,Host, and Popularity-based methodologies. Online Learning Algorithms and SVM were the machine learning algorithms employed in these investigations.[9] offer a method for detecting malicious URLs based on dynamic URL behaviors. URL properties are extracted in this paper based on both static and dynamic behaviors. Character and semantic groups are explored, as well as the Abnormal group in websites and the Host-based group Correlated group.

C. Malicious URL Detection Tools

URL Void is a URL checking application that uses various engines and domain blacklists. Google Safe Browsing ,Norton Safe Web, and My WOT [10] are some examples of URL Void. The advantage of the Void URL tool is that it works with a variety of browsers and can support a variety of additional testing services. The biggest drawback of the Void URL tool is that the process of detecting malicious URLs is mainly reliant on a set of signatures. Unmask Para sites: Unmask Parasites is a URL checking tool that works by downloading specified links and decoding HTML codes, particularly external links, I frames, and JavaScript.This tool has the advantage of being able to detect I frames quickly and precisely. This tool, however, is only useful if the user suspects something .Dr.Web Anti-Virus Link Checker is a Chrome, Fire fox,Opera, and Internet Explorer add-on that automatically detects and scans dangerous information on download links from social networking sites such as Facebook,Vk.com, and Google+. This is a Comodo malware and security hole detection tool. Users can use this to check URLs, while webmasters can use it to schedule daily checks by down loading all the necessary sites. And execute them in sandbox browser environment other resources include: Other URL checking tools include UnShorten. it, Virus Total, Norton Safe Web,Site Advisor(by Mc Afee),Sucuri ,Browser Defender, Online Link Scan, and Google Safe Browsing Diagnostic, in addition to the aforementioned standard tools. The bulk of existing harmful URL detection solutions are signature-based URL detection systems ,according to the analysis and evaluation of malicious URL detection technologies described above .As a result, these tools 'effectiveness is limited [11].

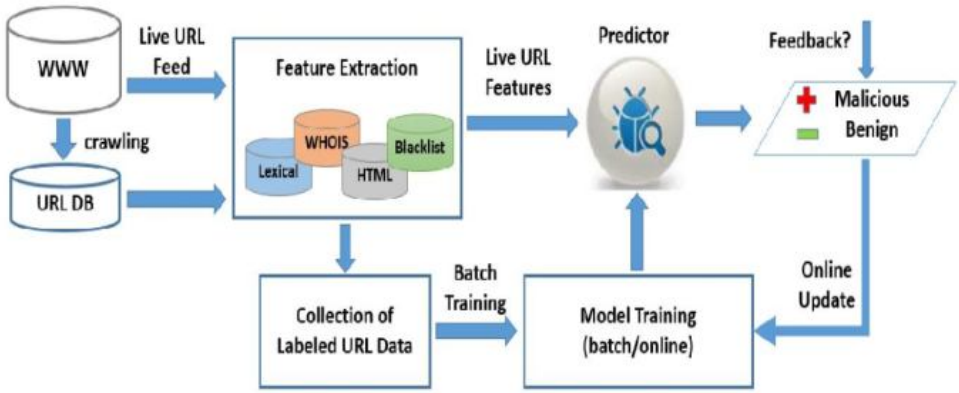
III. PROBLEM STATEMENT

Most of the existing methods to detect malicious web site are based on their core techniques for a well-known attack. But the attacker invents changes in the existing approach and introduces new techniques to be embedded in webpage. The existing approach relies on the fixed set of features, but the attacker makes changes in the existing features and also introduces new features. As a result, the detection methods are not able to detect new attacks .The various techniques like signature based, features based and behavior-based approaches to detect malicious website and contents are facing these limitations due to sophisticated invasions. Due to the limitations the various existing features are not sufficient to detect malicious websites. A part from that none of the feature collection techniques are able to collect the emerging features. The existing detection methods suffer a lot from the true and false negatives. So there is need a for new approach to overcome all these eliminations [12].

IV. MALCIOUS WEBSITE DETECTION USING URL BASED SOLUTION

A. The Model

The suggested machine learning-based dangerous URL detection system is shown in Figure 1. The machine learning-based malicious URL detection approach has two stages: training and detection. To detect malicious URLs, both harmful and clean URLs must be collected during the training stage. After that ,all of the malicious and clean URLs are accurately identified, and attribute extraction begins. These characteristics will be the most reliable indicators of which URLs are safe and which are dangerous. In this study, the specifics of these characteristics will be discussed in depth. Finally, this dataset is split into two sections: training data for machine learning algorithms and testing data for the testing procedure. If the machine learning model's classification performance is good (high classification accuracy), Step of detection: Each input URL undergoes a phase of detection.



The URL will first be subjected to an attribute extraction method. The classifier then uses these attributes to determine whether the URL is clean or malicious [7].

B.URL Attribute Extraction and Selection

The following primary attribute groups for malicious URL detection. URL length, primary domain length, maximum token domain length, path average length, and average token length in domain are examples of lexical characteristics. Host-based Features: These attributes are derived from the URLs' host characteristics. These attributes reveal the location of malicious servers, their identities, and the extent to which certain host-based factors contribute to the URL's harmful level.

D. Machine Learning Algorithm Selection

Machine learning methods have been extensively investigated and deployed in the detection of dangerous URLs12].Two widely utilized Supervised machine learning techniques, RF and SVM[13], are used in this research. Machine learning techniques are the final piece of the jigsawin this study's suggested malicious URL detection system. Those algorithms are well suited to maximizing the use of our new malicious URL detection tools .Machine learning algorithms have previously been well researched in the literature. SVM and RF are used as examples in this paper to demonstrate the overall detection system's high performance, but they are not the major focus. Other algorithms such as Nave Bayes, Decision Trees, K Nearest Neighbors, and Neural Networks are urged to be implemented by readers. Different parameter adjustments are used to investigate the effectiveness of using these two methods [14].

V. PROPOSEDWORK

Experiment dataset: The experimental dataset for the harmful URL detection [15] model contains 470.000URLs gathered from [16], with roughly 70.000 malicious URLs and safe URLs. The Virus Total programme is used to verify the labels of all of these URLs.CSV format is used to store the entire dataset. Each URL sample has a label that reads "bad" if it's malicious and"good"if it's not. The following Phishtank is a website that allows users to share phishing URLs. URLs that appear to be suspicious can be forwarded to Phishtank for verification. Phishtank's data is updated every hour. URL haus is a project by abuse that aims to share dangerous URLs that are used to distribute malicious software. Alexa[18] is a database that ranks all websites based on their usefulness. Harmful and Non-Malicious URLs[19] this could be a data source with over 400,000 URLs classified as malicious or non-malicious. According to this database, 82 percent of the URL sare safe, while the remaining18 percent are harmful.

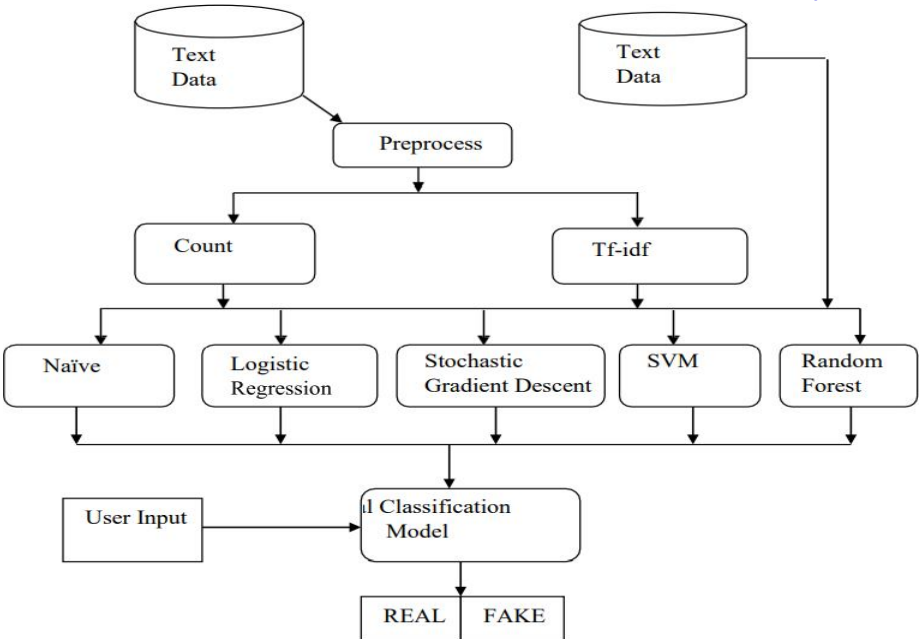
Advantages of the proposed work

Accuracy: the percentage of correct decisions among all testing samples where: TP- True positive is the number of malicious URLs correctly labeled.

FN - False negative is the number of malicious URLs misclassified as safe; TN-True negative is the number of safe URL correctly labeled; FP – False positive is the number of safe URLs mis classified as malicious. Precision is the percentage of correctly categorized malicious URLs (TP) out of all malicious URLs labeled by the classifier (TP+FP).

VI. SYSTEM ARCHITECTURE

Following is the architecture of the proposed system. It has three modules which are used to classify a given article into malicious website or real website. The model has built all the classifiers for predicting malicious websites. The extracted features are fed into different classifiers. They have used Naive-bayes, Logistic Regression, Linear SVM, Stochastic gradient decent and Random Forest classifiers from sk learn.Once fitting the model, compared the f1 score and checked the confusion matrix [20].



VII. IMPLEMENTATION

Modules: We have designed this project with the following modules.

- Data cleaning
- Features election
- Model selection and prediction
- Save the model

Modules Description:

1. **Data Cleaning** Data cleaning is the process of ensuring data is correct, consistent and usable. You can clean data by identifying errors or corruptions, correcting or deleting them, or manually processing data as needed to prevent the same errors from occurring.
2. **Feature selection** Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.
3. **Model Selection** The best fit algorithm is found out for our dataset for classifying malicious and legitimate URL. Model selection is the process of choosing one among many candidate models for a predictive modeling problem. There may be many competing concerns when performing model selection beyond model performance, such as complexity, maintainability ,and available resources.
4. **Save The Model** Once we select the model for prediction then selected trained model is saved in the pickle format for further prediction. This model will be used for the detecting the user input URL as malicious or not.

VIII. METHODOLOGY

Random Forrest Random Forests are a machine learning method of classification that work by building several decision trees while training the model. It is a kind of additive model that makes predictions from a combination of decisions from base models. Decision trees have huge depth and tend to over fit results. Random forest utilizes multiple decision trees to average out the results.

Pseudo code

1. Select randomly M features from the feature set.
2. For each x in M

a. Calculate the Information Gain Gain(t,x)=E(t)-E(t,x)E(t)=∑i=1-Pi log2Pi

E(t,x)=∑ c€xP(c)E(c)

b. Select the noded which has the highest information gain.

c. Split the node into sub-nodes.d. Repeat steps a,b and c to construct the tree until reach minimum number of samples.

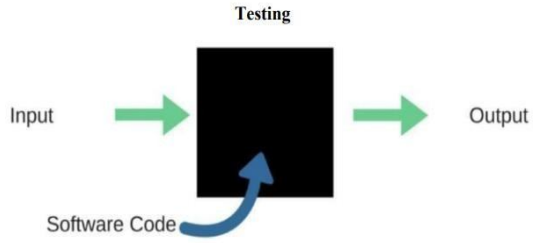
3. Repeat steps 1 and 2 for N times to build forest of N trees.

IX. SYSTEMTESTING

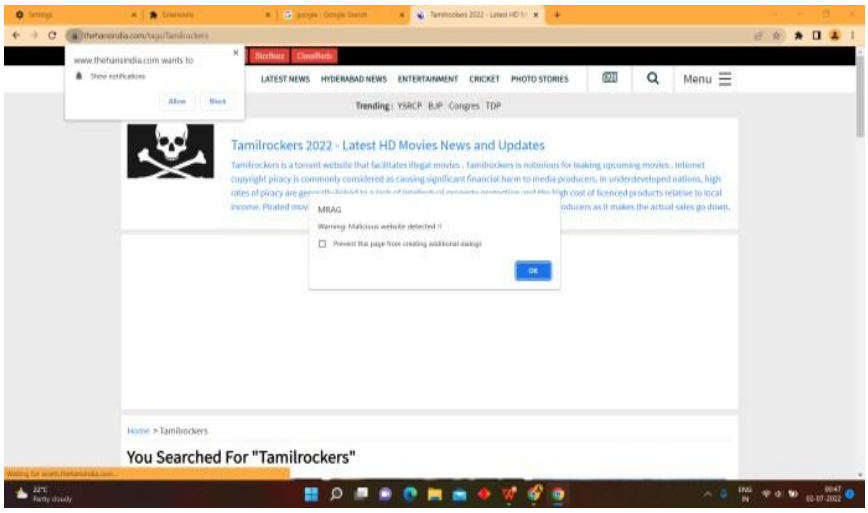
Software testing is conducted to provide Stakeholders with information about the quality of the software productor service under test. Software testing can also provide an objective, independent view of the software to allow the business to appreciate and understand the risks of software implementation. It involves the execution of a software component to system component to evaluate one or more properties of interest (a program or application),with the intent of finding software bugs (errors or other defects),and verifying that the software product is fit for use.

TYPES OF TESTS

Black Box Testing Blackbox testing is testing the functionality of an application without knowing the details of it simplementation including internal program structure, data structures,etc. Test cases for Blackbox testing are created based on the requirement specifications. Therefore, it is also called as specification-based testing. The following figure represents the Black boxtesting. When applied to Machine Learning models, Black boxtesting would mean testing Machine Learning models without knowing the internal details such as features of the Machine Learning model, and the algorithm used to create the model etc. These challenges however, is to identify the test oracle which could verify the test outcome against the expected values known before hand. This is discussed in the following section.



Unit testing For unit testing the different modules of the project were tested for their functionality. The model was trained using the same set of features using random forest algorithms. Each of the extracted features were used in all of the classifiers. Once fitting the model, compared the f1 score and checked the confusion matrix. After fitting all the classifiers,best performing models were selected as candidate models for malicious website classification



Test Results: All of the above-mentioned test cases were successful. There were no faults found.

Results of the test: All of the above-mentioned test cases were successful. There were no faults found.

TESTID	I
TESTSENARIO	Giving non maliciouswebsite
TESTINPUT	https://www.bnmit.org/
EXPECTEDOUTPUT	Websiteis SAFE
ACTUAL OUTPUT	Warning: Maliciouswebsitedetected
PASS/FAIL	PASS

X. RESULTS

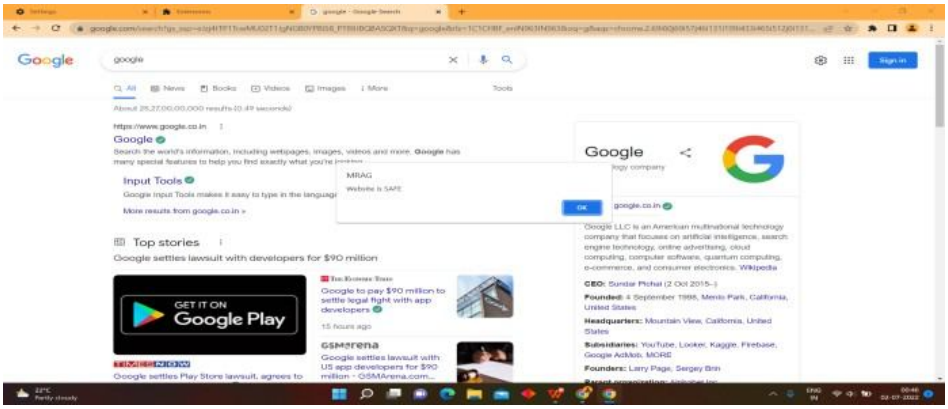


FIG1 Detecting Safe Website

If the entered website URL is correct then website is SAFE will be appeared. **FIGURE2** Detection Malicious Website If the entered URL website is fake then Malicious website page will be appeared.

CONCLUSION

A machine learning-based approach for detecting malicious URLs is provided in this paper. The efficiency of the proposed extracted qualities has been demonstrated by the empirical data in Tables V and VI. We do not use special qualities in this work, nor do we attempt to construct large datasets to improve the system's accuracy, as many other traditional articles do. The processing time and accuracy of the system are determined by the combination of easy-to-calculate qualities and big data processing technologies to ensure the balance of the two elements. The findings of this study can be used to develop and deploy information security technologies and systems. The findings of this article were utilized to create a free tool[20] that can detect rogue URL son the web.

REFERENCES

[1]. Sahoo, C. Liu, S.C.H. Hoi, Malicious URL Detection using Machine Learning: A survey".CoRR,abs/1701.07179,2017.
[2]. M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection;:aliterature survey,"IEEE Communications Surveys & Tutorials, vol. 15,no. 4, pp. 2091–2121, 2013.
[3]. M. Cova, C. Kruegel, and G. Vigna, "Detectionand analysis of drive by download attacks and malicious javascript code," in Proceedings of the19th international conference on Worldwide web. ACM, 2010, pp. 281– 290.
[4]. R. Heartfield and G. Loukas, "A taxonomy ofattacks and a survey of defence mechanisms for semantic social engineering attacks,"ACM Computing Surveys (CSUR), vol. 48, no. 3, p.37, 2015.
[5]. Internet Security Threat Report(ISTR)2019–Symantec. https://www.symantec.com/content/dam/symantec/docs/reports/istr-24- 2019-en.pdf [Lastaccessed10/2019].
[6]. S. Sheng, B. Wardman, G. Warner, L. F.Cranor,J.Hong,andC.Zhang,"An empirical analysis of phishing blacklists,"inP roceedings of Sixth ConferenceonEmailand Anti-Spam (CEAS),2009.
[7]. C. Seifert, I. Welch, and P. Komisarczuk,"Identification of malicious webpages with staticheuristics,"in Telecommunication Networks and Applications Conference, 2008. ATNAC2008. Australasian. IEEE, 2008, pp. 91–96.
[8]. S.Sinha, M.Bailey,andF.Jahanian,"Shades of grey: On the effectiveness of reputation-based" black lists","in Malicious and Unwanted Software, 2008.MALWARE2008 .3rd International Conferenceon.IEEE,2008,pp. 57–64.
[9]. J. Ma, L. K. Saul, S. Savage, and G. M.Voelker, "Identifying suspicious urls: an application of large-scale online learning,"in Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009, pp. 681–688.
[10]. B.Eshete,A.Villafiorita,andK.Weldemariam,"Binspect:Holisticanalysis and detection of malicious webpages,"in Security and Privacy in Communication Networks.Springer,2013, pp. 149–166.
[11]. S. Purkait, "Phishing counter measures and their effectiveness–literature review,"Information Management & Computer Security,vol.20,no.5,pp.382–420, 2012.
[12]. Y.Tao,"Suspiciousurland device detection by logmining,"Ph.D. dissertation, Applied Sciences: School of Computing Science, 2014.
[13]. G.Canfora,E.Medvet,F.Mercaldo,andC.A.Visaggio,"Detectionof malicious webpages using system calls sequences,"in Availability, Reliability, and Security in Information Systems. Springer,2014, pp. 226–238.
[14]. LeoBreiman.:RandomForests.MachineLearning45 (1), pp. 5-32, (2001).
[15]. Thomas G. Dietterich. Ensemble Methods inMachineLearning.InternationalWorkshoponMultiple Classifier Systems,

pp 1-15, Cagliari,Italy,2000.
[16]. DeveloperInformation.https://www.phishtank.com/developer_info.php. [Last accessed11/2019].
[17]. URLhaus DatabaseDump.https://urlhaus.abuse.ch/downloads/csv/. [Ngàytruynhập 11/2019].
[18]. DatasetURL.http://downloads.majestic.com/majestic_million.csv.[Lastaccessed 10/2019].
[19]. Malicious_n_Non-MaliciousURL.https://www.kaggle.com/antonyj453/urldataset#data.csv.[Lastaccessed11/2019].
[20]. H. Allcott and M. Gentzkow, Social media andmaliciouswebsiteinthe2016election,inNationalBureauofEconomicResearch,WorkingPaper 23089, January 2017.
[21]. S. Feng, R. Banerjee, and Y. Choi, Syntacticstylometry for deception detection, Proceedingsof the 50th Annual Meeting of the AssociationforComputationalLinguistics:ShortPapers-Volume2,ser.ACL'12.Stroudsburg,PA,USA:AssociationforComputational
[22]. Gupta,P.Kumaraguru,C.Castillo,and
[23]. P. Meier, Tweetcred: Realtime credibilityassessmentofcontentontwitter,inInternationalConferenceonSocialInformatics.
[24]. V.W.FengandG.Hirst,Detectingdeceptiveopinionswithprofilecompatibility, in Proceedings of the SixthInternational Joint Conference on NaturalLanguageProcessing, 2013, pp. 338–346.
[25]. D. M.Markowitz andJ.T.Hancock,Linguistic traces of a scientific fraud: Thecase of diederik stapel, PloS one, vol. 9,no.8, p. e105937, 2014.
[26]. D. Pisarevskaya, Deception detection inwebsite reports in the Russian language:Lexics and discourse, in Proceedings ofthe2017EMNLPWorkshop:NaturalLanguageProcessingmeetsJournalism,2017.
[27]. V. L. Rubin, Y. Chen, and N. J. Conroy,Deceptiondetectionforwebsite:Threetypes of fakes, in Proceedings of the 78thASIS&TAnnualMeeting:InformationScience with Impact: Research in and forthe Community, 2015
[28]. K.Shu,A.Sliva,S.Wang,J.Tang,and
[29]. H.Liu,Maliciouswebsitedetectiononsocial media: A data mining perspective.NaturalLanguageProcessingmeetsJournalism,2016.
[30]. SeyedAliMiraftabzadeh,PaulRad,Kim-KwangRaymondCho,Detectingdeceptiveopinionswithprofilecompatibility,2014