School of Computer Science and Engineering

Phishing Website Detection using Machine Learning Algorithms

J component Report
Review-3

ISM

Slot: F2

Gaurav Srivastava(19BCE2358)

B. Tech Computer Science and Engineering

Under Guidance of:
R. Vidhya
Associate Professor, SCOPE
VIT

Abstract

Aim

To detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

Objective

- To overcome the drawbacks of blacklist and heuristics-based method
- Focus on machine learning techniques.

Motivation

- Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Main aim of the attacker is to steal banks account credentials.
- Phishing attacks are becoming successful because lack of user awareness.
- The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as “blacklist" method.
- To overcome the drawbacks of blacklist and heuristics-based method, many security researchers now focused on machine learning techniques.
- Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero- hour phishing websites.

Introduction

In this digital day and electronic world, Internet plays a vital role in day-to-day activities like communication, business, transactions, personal needs, marketing, e-commerce etc. Internet is a multifaceted facility which helps in completing many tasks readily and conveniently within few seconds. Almost everything is presently accessible over web in this period of progression of advances. Thus, increasing usage of internet leads to cybercrime and other malware activities. The information divulged in online leaves digital imprint and if it happens to drop into the wrong hands, it will result in data theft, identity theft and monetary loss. Cybercrime includes many kinds of security issues over the internet and one of the most threatening problems is Phishing. Phishingis

a fraudulent technique achieved by phishing web page. Phishing uses e-mails and websites, which are intended to look like from trusted organization, to hoodwink clients into unveiling their own or money related data. The threatening party then use these data for criminal purposes, such as, identity or data theft and extortion. Clients are deceived into revealing their data either by giving touchy data through a web shape or downloading and introducing unfriendly codes, which seek clients' PCs or checking clients' online actions to get data. Luring Internet users by making them click on rogue links that seem trustworthy is an easy task because of widespread credulity and unawareness. It is important to prevent user's confidential data from unauthorized access. The procedure for the most part includes sending messages that then cause the beneficiary to either visit a deceitful site and enter their data or to visit an authentic site through a phishing intermediary attack or using spoofed website, which then gathers the details of user leads to several loss. The Phishing problem needs to be mitigated by anti-Phishing approaches. This research provides a solution that helps in detecting and preventing Phishing attacks using the features of phishing URLs and an automated real-time detection of phishing websites by machine learning approach.

Advantages and Disadvantages of the previous used methods

Blacklists: Blacklists hold URLs (or parts thereof) that refer to sites that are considered malicious. Whenever a browser loads a page, it queries the blacklist to determine whether the currently visited URL is on this list. If so, appropriate countermeasures can be taken. Otherwise, the page is considered legitimate. The blacklist can be stored locally at the client or hosted at a central server.

Advantages:

Obviously, an important factor for the effectiveness of a blacklist is its coverage. The coverage indicates how many phishing pages on the Internet are included in the list. Another factor is the quality of the list. The quality indicates how many non-phishing sites are incorrectly included into the list. For each incorrect entry, the user experiences a false warning when she visits a legitimate site, undermining her trust in the usefulness and correctness of the solution.

Finally, the last factor that determines the effectiveness of a blacklist-based solution is the time it takes until a phishing site is included. This is because many phishing pages are short-lived and most of the damage is done in the time span between going online and vanishing. Even when a blacklist contains many entries, it is not effective when it takes too long until new information is included or reaches the clients.

Disadvantages:

The overall technique to identify phishing sites by refreshing boycotted URLs, Internet Protocol (IP) to the antivirus information base which is otherwise called "boycott" strategy. To dodge boycotts, assailants utilizes innovative procedures to trick clients by adjusting the URL to seem real through muddling and numerous other basic methods including: quick motion, in which intermediaries are naturally produced to have the page; algorithmic age of new URLs; and so forth. Significant disadvantage of this strategy is that it can't recognize Zero-hour phishing attack.

Test Data and Statistics

For our study, a large number of phishing pages were necessary. We concatenated three databases from Kaggle and merged it into one. The information collected by this method is freely available and the amount of reported phishing sites is very large.

Table: (a) Domains that host phishing sites. (b) Popular phishing targets.

(a)

No domain (numerical)	3,864
.com	1,286
.biz	1,164
.net	469
.info	432
.ws	309
.jp	307
.bz	256
.nz	228
.org	156
.de	111
.ru	106
.us	105

(b)

paypal	1,301
53.com	940
ebay	807
bankofamerica	581
barclays	514
volksbank	471
sparkasse	273
openplan	182
Total	5,069

LITERATURE SURVEY

Title	Author	Journal	Methodology	Pros-Cons	Challenges
1. Phishing E-mail Detection Based on Structural Properties	Chandrasekaran, Madhusudhanan, Krishnan Narayanan,and Shambhu Upadhyaya.	NYS cyber security conference. Vol. 3	The proposed approach explains to find phishing through appropriate identification and usage of structural properties of email. The experiment is done by SVM and classification technique to classify phishing e-	It uses proper identification and use of structural properties as it's advantage. The disadvantage being it is low in efficiency.	The technique used in this classification method is not large enough and it uses only one approach to identify phishing e-mails, which is low in efficiency and scalability. This is purely based on structural properties of e-mail and it has to extend more structural.or

			mails.		content properties to reduce error results.
2. Discovering Phishing Target Based on Semantic Link Network	Wenyin, L., Fang, N., Quan, X., Qiu, B., & Liu, G.	Future Generation Computer Systems 26, no. 3 (2010): 381-388.	The.paper proposes a novel approach.to discover phishing website.by calculating association relation. among webpages. that include.malicious webpages and its associated webpages.to measure,the combination of link relation,search relation, and text relation.	Advantage is just that it calculates association relation with websites. While the disadvantage being that it is very time consuming	The demerits in this approach are more kind of association has to be done, similarities between visual, layout and domain have to be related. This method is considered as a time-consuming approach and also various sub-relations in the combined association relations be studied.
3. Evolving Fuzzy Neural Network for Phishing Emails Detection	Almomani, Ammar, et al	Journal of Computer Science 8. 7 (2012): 1099.	It deals with zero-day phishing email. It differentiates phishing email and ham email in online mode. It is adopted on feature fetching, rank	Pro is that it differentiates between phishing and ham email. Con is that it is not much dynamic	This technique does not have more dynamic system, so it is less.in performance to produce accurate results.

			fetching and grouping similar features of email.	system.	
4. Intelligent Phishing Website Detection and Prevention System by Using Link Guard Algorithm	CV, U. Naresh1 U. Vidya Sagar, and Madhusudan Reddy.	_	It proposed a system using link guard algorithm which works for hyperlinks. The algorithm performs certain tests like comparison of the DNS of actual and visual links, checks dotted decimal of IP address, checks encoded links and pattern matching.	Pro is that the algorithm works for hyperlinks too. Con is that it sometimes gives false conclusions.	The drawbacks of this system is, it produce the false positive results if any genuine site has IP address instead of domain name, and it considers some phishing site as normal one if the user does not visit the original site. This results in false negative conclusions.
5. Said Afroz, Rachel Greenstadt – Phishzoo Approach	Preethi, V., and G. Velmayil.	International Journal of Engineering and Techniques 2.5 (2016): 107-115.	The algorithm detects current phishing sites by matching their content with genuine site. This will match images, contents and the structure of website with trusted one in order to avoid phishing.	Pro is that it matches the content with actual and genuine websites. Con is that it is less robust for detect phishing attacks.	Drawbacks of this algorithm is, it requires matching image site, and it is less robust for detecting phishing attacks.

6. Phishing website detection using machine learning.	Purvi Pujara, MB Chaudhari	International Journal of Scientific Research in Computer Science, Engineering and Information Technology 3 (7), 395-399, 2018	Obtain sensitive information such as username, password, bank account details, and credit card details for malicious use.		Several anti-phishing techniques are there such as blacklist, heuristic, visual similarity and machine learning. From this, blacklist approach is commonly used because it is easy to use and implement but it fails to detect new phishing attacks
7. Phishing website detection based on supervised machine learning with wrapper features selection	Waleed Ali	International Journal of Advanced Computer Science and Applications 8 (9), 72-78, 2017	This paper presents a methodology for phishing website detection based on machine learning classifiers with a wrapper features selection method.	The experimental results demonstrated that the performance of the machine learning classifiers was improved by using the wrapper-based features selection. Moreover, the machine learning classifiers with the wrapper-based features selection outperformed the machine learning classifiers with other	Detecting phishing websites is a challenging task, as most of these techniques are not able to make an accurate decision dynamically as to whether the new website is phishing or legitimate.

				features selection methods	
8. Intelligent phishing website detection using random forest classifier	Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J Chaudhery	2017 International conference on electrical and computing technologies and applications (ICECTA), 1-5, 2017	In this paper, an intelligent system to detect phishing attacks is presented. We used different data mining techniques to decide categories of websites: legitimate or phishing. Different classifiers were used in order to construct accurate intelligent system for phishing website detection. Classification accuracy, area under receiver operating characteristic (ROC) curves (AUC) and F-measure is used to evaluate the performance of the data mining techniques	Results showed that Random Forest has outperformed best among the classification methods by achieving the highest accuracy 97.36%. Random forest runtimes are quite fast, and it can deal with different websites for phishing detection.	–
9. Improving spoofed website	Ekta Gandotra, Deepak	Cybernetics and	This paper brings out a diverse set of	The experimental results	Fake webpages/websites are created by cyber attackers who either

detection using machine learning	Gupta	Systems 52 (2), 169-190, 2021	robust features categorized into the three categories, i.e., webpage, URL and HTML based features. The features under these categories are firstly used individually to classify webpages. Thereafter, a technique is proposed where the integration of all the features is used for classification purpose	demonstrate that the features under URL based category are most effective in classifying the webpages. Further, there occurs a significant improvement in classification accuracy using proposed approach and random forest turns out to be the best classifier offering the accuracy of 99.5% with FPR and FNR as 0.006 and 0.005 respectively.	try to advertise their products, attempt to transmit malware to the target device, or steal victims' login credentials.
10. Phishing web site detection using diverse machine learning algorithms	Ammara Zamir, Hikmat Ullah Khan, Tassawar Iqbal, Nazish Yousaf, Farah Aslam, Almas Anjum,	The Electronic Library, 2020	Features of phishing data set are analysed by using feature selection techniques including information gain, gain ratio, Relief-F and recursive feature elimination	_	It's implementation has not been done yet.

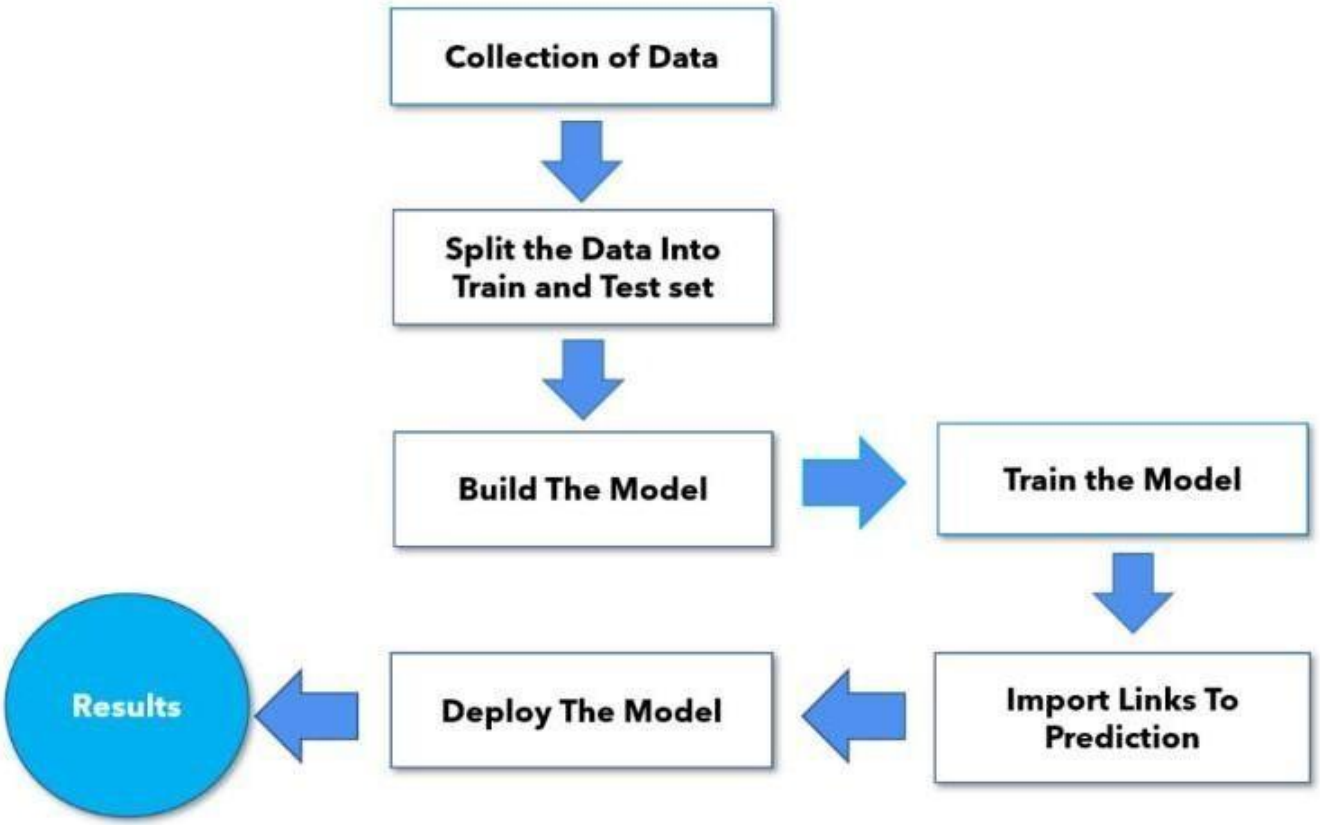
	Maryam Hamdani		(RFE) for feature selection. Two features are proposed combining the strongest and weakest attributes. Principal component analysis with diverse machine learning algorithms including (random forest [RF], neural network [NN], bagging, support vector machine, Naïve Bayes and k-nearest neighbour) is applied on proposed and remaining features. Afterwards, two stacking models: Stacking1 (RF + NN + Bagging) and Stacking2 (kNN + RF + Bagging) are applied by combining highest scoring classifiers to improve the classification		
11. Phishing websites detection using	Kulkarni, Arun D., and Leonard	_	Using data sets to compare with some preset classifiers to	Futuristic approach towards Phishing	Very few data sets are available containing Phishing URLs in Public. Due to this limitation,

machine learning.	L. Brown III.		determine if it's a phishing URL or not. Specifically, it used the decision tree, Naïve Bayes' classifier, SVM, and the Neural Network to classify the URLs in the data set, and then we compared the results using confusion matrices.	detection. Limitations include, limited data-sets and all features are discrete.	extensive studies and surveys are needed to evaluate the effectiveness of ML detection based on the existing data sets.
12. Towards detection of phishing websites on client-side using machine learning based approach	Jain, Ankit Kumar, and Brij B. Gupta	Telecommunication Systems 68.4 (2018): 687-700.	Get the source code and URL for a webpage. Split each part of the code into separate classes. Split it into the feature vector and make the training and testing sample for ML. Then run it through the RF classifier to check if it's a legitimate website or a phishing website.	Some pros of this approach are language independence, low response time, third-party independence, compromised domain detection, and client-side application.	Can detect phishing within only webpage URLs and source code. Works with only HTML written code only and cannot detect phishing within non-HTML sources. Similarly, phishing websites in the mobile environment is also a challenge.
13. Detection of phishing URLs using machine learning techniques.	James, Joby, L. Sandhya, and Ciza Thomas	2013 international conference on control communication and computing (ICCC). IEEE, 2013.	The first step is the collection of phishing URLs. Then using the host-based, popularity-based, and lexical-based feature extractions a	The approach of using lexical features points to increased efficiency.	This approach makes use of blacklists to facilitate phishing detection. The problem of such an approach is the need to construct blacklists in advance which in turn gives

			classifier is extracted. Then the classifier is implemented within ML to determine the legitimacy of the URL.	But the main con of this approach is stale blacklists.	rise to the primary problem which is these lists becoming stale.
14. Detecting phishing websites using machine learning	Alswailem, Amani, et al	2019 2nd International Conference on Computer Applications & Information Security (ICCAIS). IEEE, 2019.	In this approach, firstly you collect the data set and the websites. Extract the features from the websites to be processed through the ML algorithm. The algorithm studies all extracted features and sorts them while removing irrelevant ones. Then we use the RF algorithm to test the legitimacy of the website.	Reduced time for computation and provides high efficiency in determining the legitimacy of a website. It also provides a high accuracy in detection.	Much of a theoretical concept.
15. Phishing Detection Using Machine Learning Techniques.	Shahrivari, Vahid, Mohammad Mahdi Darabi, and Mohammad Izadi	arXiv preprint arXiv:2009.11116 (2020).	For evaluating phishing classification performance this approach uses accuracy, recall, precision, F1 score, test time, and train time of classifiers to generate custom formula which	Easy to understand and visulaize algorithm. Slightly slower due to using AdaBoost in the algorith.	One of the main challenges of this approach is the scarcity of data-sets. Although many scientific papers about phishing detection have been published, they have not provided the dataset on which they used in their research. Moreover, another factor that

			determines the accuracy of the test data.		hinders finding a desirable dataset is the lack of a standard feature set to record characteristics of a phishing website.

Proposed Methods



Data collection

Collecting data for training the ML model is the basic step in the machine learning pipeline. The predictions made by ML systems can only be as good as the data on which they have been trained. Following are some of the problems that can arise in data collection:

Inaccurate data. The collected data could be unrelated to the problem statement.

Missing data. Sub-data could be missing. That could take the form of empty values in columns or missing images

for some class of prediction.
Data imbalance. Some classes or categories in the data may have a disproportionately high or low number of corresponding samples. As a result, they risk being under-represented in the model.
Data bias. Depending on how the data, subjects and labels themselves are chosen, the model could propagate inherent biases on gender, politics, age or region, for example. Data bias is difficult to detect and remove

Train-Test Split Evaluation

The train-test split is a technique for evaluating the performance of a machine learning algorithm.
It can be used for classification or regression problems and can be used for any supervised learning algorithm.
The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

Train Dataset: Used to fit the machine learning model.

Test Dataset: Used to evaluate the fit machine learning model.

The objective is to estimate the performance of the machine learning model on new data: data not used to train the model. This is how we expect to use the model in practice. Namely, to fit it on available data with known inputs and outputs, then make predictions on new examples in the future where we do not have the expected output or target values.

Model Building

We finally now use the prepared data for model building. Depending on the data type (qualitative or quantitative) of the target variable (commonly referred to as the Y variable) we are either going to be building a classification model. We will be using logistic regression, MultinomialNB and Random Forest.

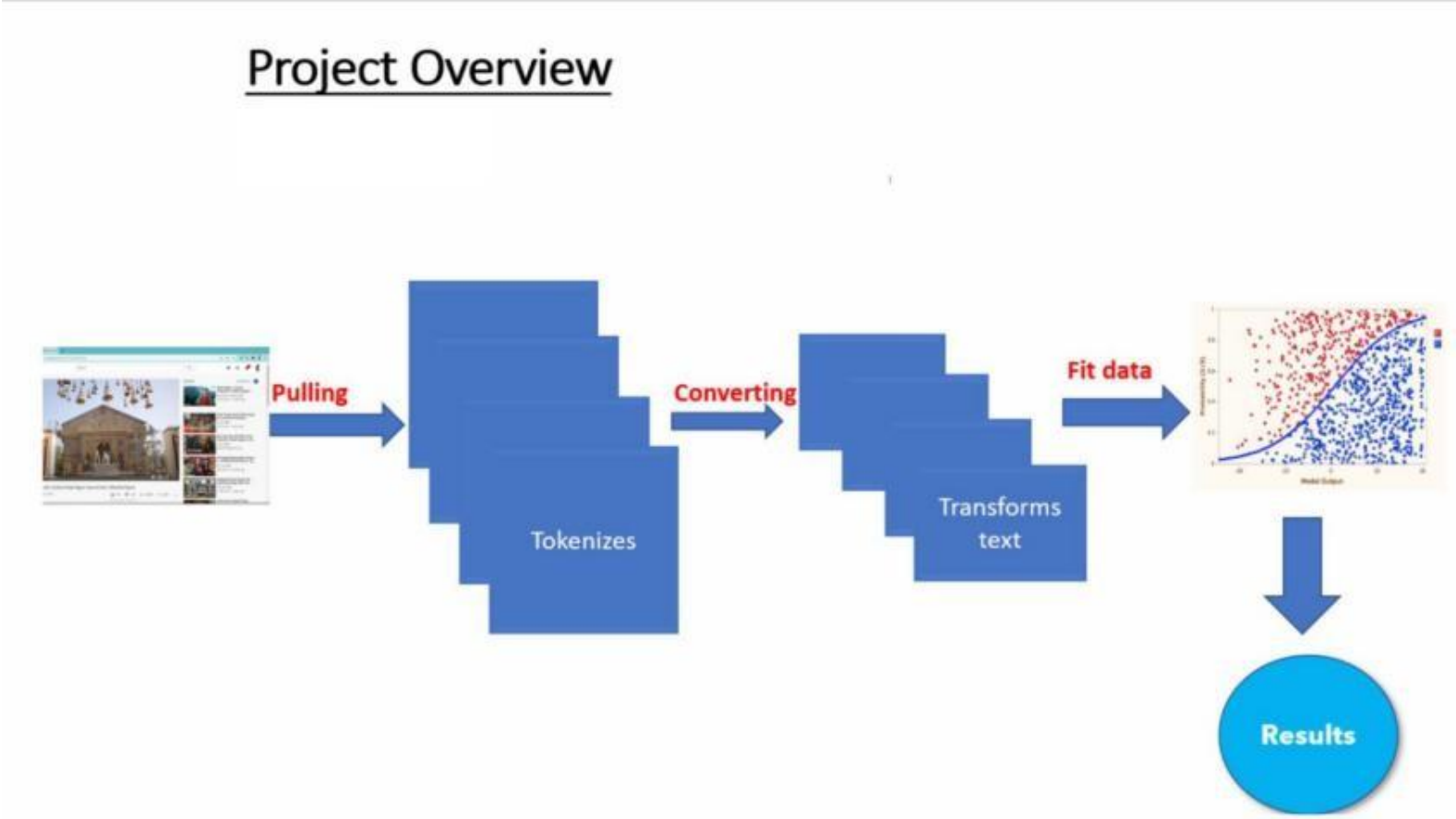
Model training

It consists of the sample output data and the corresponding sets of input data that have an influence on the output. The training model is used to run the input data through the algorithm to correlate the processed output against the

sample output. The result from this correlation is used to modify the model.

Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data. It is one of the last stages in the machine learning life cycle and can be one of the most cumbersome.

WORKING OF THE PROJECT



Detailed Description of Methodology

MACHINE LEARNING ALGORITHMS

- Logistic regression

  Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation.

- Random Forest Algorithm

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

- Multinomial Naive Bayes

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output. The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts.

DATASET

We combined various datasets from Kaggle and concatenate them into one. The final file name after concatenation is phishing_classifier_url.csv. The dataset furthermore contains 2 attributes which are:

- URL: this contains a list of many URLs which are either phishing sites or are not phishing sites.
- LABEL: this comprises of a binary entry i.e. bad or good.

RESULT AND ANALYSIS OF PROPOSED METHOD

We used the same legitimate URL “google.com” on each algorithm and found the following accuracy:

1. Random Forest Classifier:

Accuracy: 70%

2. Logistic regression classifier:

```
CONFUSION MATRIX
: <matplotlib.axes._subplots.AxesSubplot at 0x1ec84c387c8>
```

3. MultinomialNB classifier:

```
CONFUSION MATRIX
Out[51]: <matplotlib.axes._subplots.AxesSubplot at 0x1ec84f9c908>
```

Table below shows the results of classifiers used for the classification process in python. From the table it is shown that the classifier **Logistic Regression** produce the best result.

Classifier	Accuracy
Random Forest	70%
Logistic regression	96.36%
MultinomialNB	95.78%

CONCLUSION

This project intends to upgrade recognition technique to recognize phishing sites utilizing machine learning innovation. From our experiment we found logistic regression has the highest accuracy rate as 96.3% as compared to random forest and logistic regression classifier. Likewise result shows that classifiers give better execution when we utilized more information as preparing information. Hence the users can get rid of phishing sites and be safe by avoiding them using this technique. The final confusion matrix after optimizing the logistic regression and completing the pipelining work is given below.

REFERENCES

[1] Chandrasekaran, Madhusudhanan, Krishnan Narayanan, and Shambhu Upadhyaya. "Phishing email detection based on structural properties." NYS Cyber Security Conference. 2006.

[2] Wenyin, Liu, et al. "Discovering phishing target based on semantic link network." Future Generation Computer Systems 26.3 (2010): 381- 388.

[3] Almomani, Ammar, et al. "Evolving fuzzy neural network for phishing emails detection." Journal of Computer Science 8.7 (2012): 1099.

[4] Madhuri, M., K. Yeseswini, and U. Vidya Sagar. "Intelligent phishing website detection and prevention system by using link guard algorithm." Int. J. Commun. Netw. Secur 2 (2013): 9-15.

[5] Afroz, Sadia, and Rachel Greenstadt. "Phishzoo: Detecting phishing websites by looking at them." Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on. IEEE, 2011.

[6] Purvi Pujara, MB Chaudhari. "Phishing website detection using machine learning." International Journal of Scientific Research in Computer Science, Engineering and Information Technology 3 (7), 395-399, 2018

[7] Waleed Ali."Phishing website detection based on supervised machine learning with wrapper features selection". International Journal of Advanced Computer Science and Applications 8 (9), 72-78, 2017

[8] Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J Chaudhery. "Intelligent phishing website detection using random forest classifier". 2017 International conference on electrical and computing technologies and applications (ICECTA), 1-5, 2017

[9] Ekta Gandotra, Deepak Gupta. "Improving spoofed website detection using machine learning". Cybernetics and Systems 52 (2), 169-190, 2021

[10] Ammara Zamir, Hikmat Ullah Khan, Tassawar Iqbal, Nazish Yousaf, Farah Aslam, Almas Anjum, Maryam Hamdani. "Phishing web site detection using diverse machine learning algorithms". The Electronic Library, 2020

[11] Kulkarni, Arun D., and Leonard L. Brown III. "Phishing websites detection using machine learning." (2019).

[12] Jain, Ankit Kumar, and Brij B. Gupta. "Towards detection of phishing websites on client-side using machine learning based approach." *Telecommunication Systems* 68.4 (2018): 687-700.

[13] James, Joby, L. Sandhya, and Ciza Thomas. "Detection of phishing URLs using machine learning techniques." *2013 international conference on control communication and computing (ICCC)*. IEEE, 2013.

[14] Alswailem, Amani, et al. "Detecting phishing websites using machine learning." 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS). IEEE, 2019.

[15] Shahrivari, Vahid, Mohammad Mahdi Darabi, and Mohammad Izadi. "Phishing Detection Using Machine Learning Techniques." *arXiv preprint arXiv:2009.11116* (2020).