PHISHING ATTACKS DETECTION USING MACHINE LEARNING

REVIEW 0

DOMAIN: MACHINE LEARNING

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans. Machine learning is actively being used today, perhaps in many more places than one would expect.

ABSTRACT

Phishing sites which expects to take the victims confidential data by diverting them to surf a fake website page that resembles a honest to goodness one is another type of criminal acts through the internet and its one of the especially concerns toward numerous areas including e-managing an account and retailing. Phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized, for example, fuzzy, neural system and data mining methods can be a successful mechanism in distinguishing phishing sites. We applied Random Forest (RF), one of the different types of machine learning based algorithms used for detection of Phishing websites. Finally we measured and compared the performance of the classifier in terms of accuracy.

PHISHING ATTACKS DETECTION USING MACHINE LEARNING

REVIEW 1

ABSTRACT

Phishing sites which expects to take the victims confidential data by diverting them to surf a fake website page that resembles a honest to goodness one is another type of criminal acts through the internet and its one of the especially concerns toward numerous areas including e-managing an account and retailing. Phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized, for example, fuzzy, neural system and data mining methods can be a successful mechanism in distinguishing phishing sites. We applied Random Forest (RF), one of the different types of machine learning based algorithms used for detection of Phishing websites. Finally we measured and compared the performance of the classifier in terms of accuracy.

EXISTING SYSTEM

There are countless domains where phishing attack canoccur 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. Data is collected from Phish Tank and legitimate URLs from Google.To obtain the text properties C# programming and R programming were used. 133 features were obtained from the dataset and third party service providers. CFS subset based and Consistency subset based feature selection methods used for feature selection and analyzed with WEKA tool. Naïve Bayes and Sequential

Minimal Optimization (SMO) algorithms were compared for performance evaluation and SMO is preferred by the author for phishing detection than NB.

DRAWBACKS OF EXISTING SYSTEM

- Efficiency is low. More number of repeated work.
- Zero day phising email attack.

LITERATURE SURVEY

Sl.No	Title	Author	Description	Technique
1.	Detecting Phishing Websites via Aggregation Analysis of Page Layouts	JianMao ,JingdongBian ,WenqianTian ,ShishiZhu ,TaoWei ,AiliLi ,ZhenkaiLiang/ 2018	In this paper, we aim to improve phishing detection techniques using machine learning techniques. In particular, we propose a learning-based aggregation analysis mechanism to decide page layout similarity, which is used to detect phishing pages. Our experiment results shows that our approach is accurate and effective in detecting phishing pages.	Support Vector Machine (SVM), DecisionTree (DT)
2.	Detection of Phishing Websites using Machine Learning	Atharva Deshpande , Omkar Pedamkar , Nachiket Chaudhary , Dr. Swapna Borde/ 2021	This paper surveys the features used for detection and detection techniques using machine learning. Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computers defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organizations logos and other legitimate contents.	Random forest algorithm and Decision tree

3.	A Novel Machine Learning Approach to Detect Phishing Websites	Ishant Tyagi; Jatin Shad; Shubham Sharma; Siddharth Gaur; Gagandeep Kaur/ 2018	This paper is focused on various Machine Learning algorithms aimed at predicting whether a website is phishing or legitimate. Machine learning solutions are able to detect zero hour phishing attacks and they are better at handling new types of phishing attacks, so they are preferred. In our implementation, we managed an accuracy of 98.4% in prediction a website to be phishing or legitimate.	Decision Tree, Random Forest, Gradient Boosting (GBM), Generalized Linear Model and K-Nearest-Neighbors (KNN).
4.	Phishing Detection Using Machine Learning Techniques	Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi/ 2020	One of the most successful methods for detecting these malicious activities is Machine Learning. This is because most Phishing attacks have some common characteristics which can be identified by machine learning methods. In this paper, we compared the results of multiple machine learning methods for predicting phishing websites.	LogisticRegression, DecisionTree, RandomForest, Ada-Boost, Support Vector Machine, KNN, Artificial Neural Networks, Gradient Boosting, and XGBoost
5.	Development of anti-phishing browser based on random forest and rule of extraction framework	Mohith Gowda HR, Adithya MV, Gunesh Prasad S & Vinay S/ 2020	In this paper, we propose a novel technique to identify phishing websites effortlessly on the client side by proposing a novel browser architecture. In this system, we use the rule of extraction framework to extract the properties or features of a website using the URL only. This list consists of 30 different properties of a URL, which will later be used by the Random Forest Classification machine learning model to detect the authenticity of the website.	Random Forest Classification machine learning model

PHISHING ATTACKS DETECTION USING MACHINE LEARNING

REVIEW 2

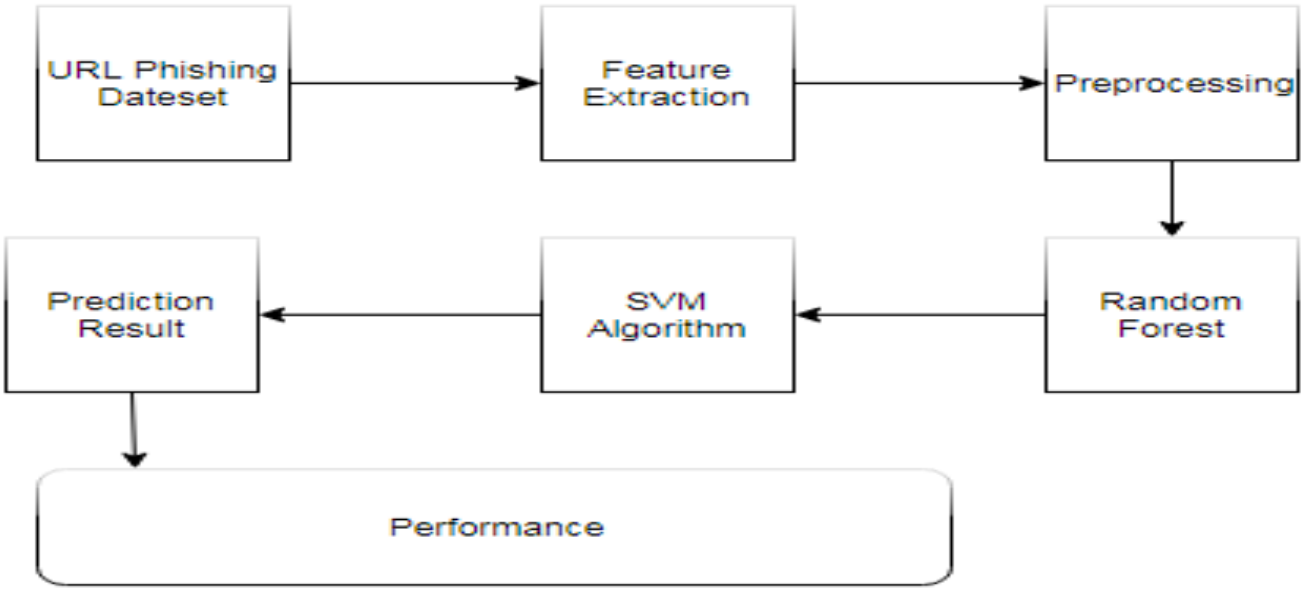
PROPOSED SYSTEM

- 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.

- This paper proposes an approach of phishing detection system to detect blacklisted URL also known as phishing websites.

- The individual 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.

SYSTEM ARCHITECTURE



ARCHITECTURE FLOW

LIST OF MODULES

- ✓ Detection Technique
- ✓ Phishing Websites Features
- ✓ Data Set

MODULE 1 : DETECTION TECHNIQUE

Detection of phishing websites has received a lot of attention recently due to their impact on users' security. Therefore, many techniques have been developed to detect phishing websites varying from communication-oriented techniques, such as authentication protocols,blacklisting, and white-listing, to content-based filtering techniques. The blacklisting and white-listing techniques have not proven though to be sufficiently efficient when used in different domains, and thus they are not commonly used. Meanwhile, the content-based phishing filters have been widely used and have proven to be of high efficiency. In light of this, researches have focused on content-based mechanism and on developing machine learning and data mining techniques based on the header and body of emails.

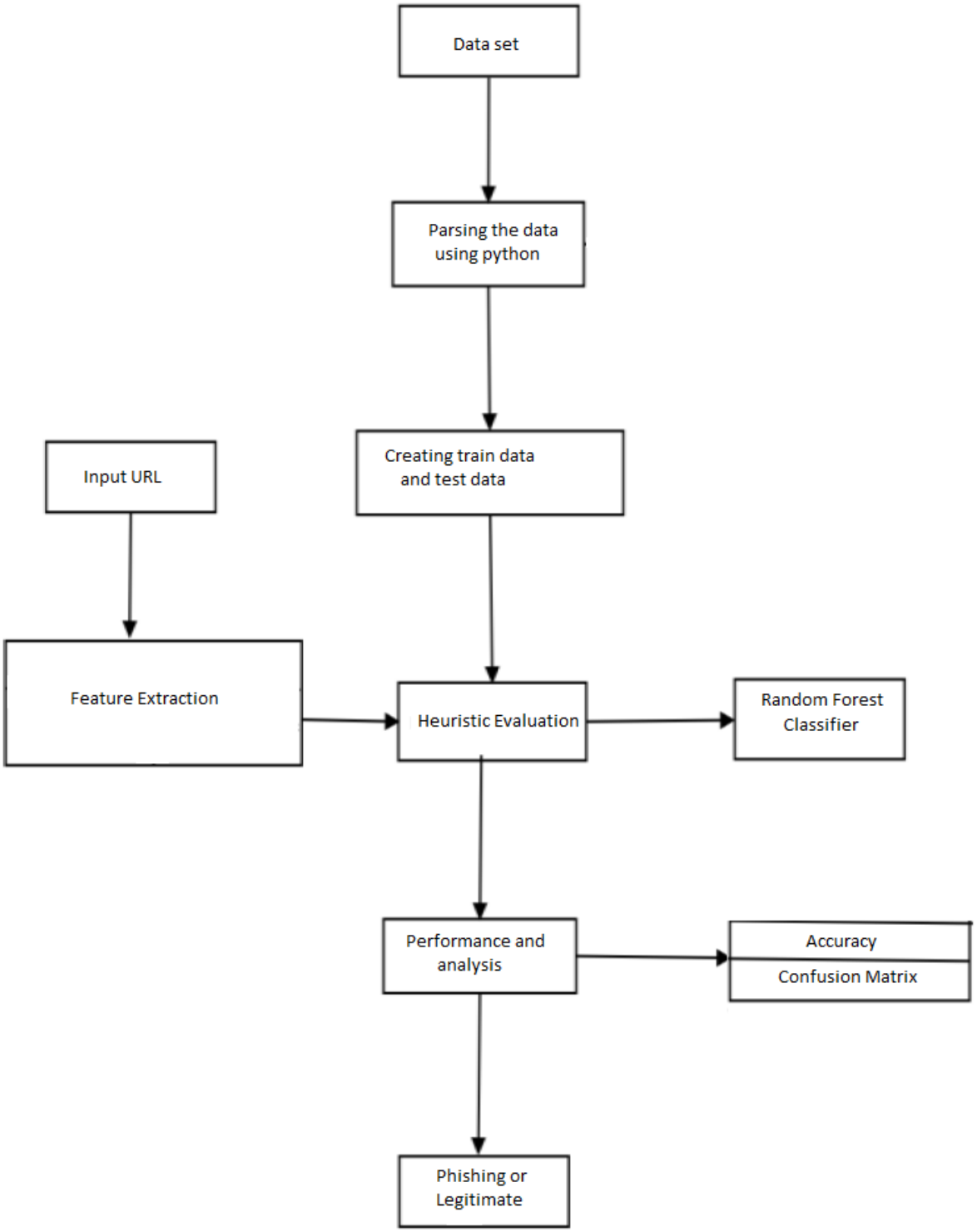
MODULE 2 : PHISHING WEBSITES FEATURES

One of the challenges faced by our research was the unavailability ofreliable training datasets. In fact, this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites using data mining techniques have been disseminated these days, no reliable training dataset has been published publically, maybe because there is no agreement in literature on the definitive features that characterize phishing websites, hence it is difficult to shape a dataset that covers all possible features. In this article, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we proposed some new features, experimentally assign new rules to some well-known features and update some other features.

MODULE 3 : DATA SET

The dataset is collected from the kaggle website which is an open source. It consists of 16 features and more than 1000 records.

DATA FLOW DIAGRAM



Data Flow Diagram