PHISHING WEB SITES FEATURES CLASSIFICATION BASED ON MACHINE LEARNING

M.Reshmi[1],S.Srinadh Raju[2], B.S.Panda[3]
[1] Student, Dept of CSE, Raghu Engineering College,A.P,India
[2]Assistant Professor, Dept of CSE, Raghu Engineering College, A.P,India
[3]Professor, Dept of CSE, Raghu Engineering College, A.P,India

ABSTRACT

Phishing is one of the most common and most dangerous attacks among cybercrimes. The main aim of these attack is to hack the user information by accessing the credentials that is used by individuals and any of the organizations. Phishing web sites contains various hints among their contents and web browser-based information. The victim's confidential data is expected by the phishing sites by deriving them to surf a phishing web sites that resembles to legitimate websites, which is one of the criminal attacks prevailing in the internet. Phishing websites is similar to cyber threat that is targeting to get all the credential-based information accessed from the credit cards and social security numbers. The purpose of this project is to perform Extreme Learning Machine (ELM) based classification. There are different types of features based on web pages. Hence, to prevent phishing attacks we must use a specific web page feature. Here, a model based on Machine Learning techniques like Naïve Bayes is used to detect phishing web pages.

KEYWORDS:*MachineLearning,Cybersecurity,PhisingDetection,Featureclassification,webHTML,Java*

I.INTRODUCTION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm inwhich it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

1.Request Clarification

2.Feasibility Study

3.Request Approval

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

FEASIBILITY STUDY :

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

1.Operational Feasibility

2.Economic Feasibility

Technical Feasibility

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility

REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, it cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

2.INPUTAND OUTPUT DESIGN:

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error is in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.

Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

OBJECTIVES:

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.
When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user.

OUTPUT DESIGN

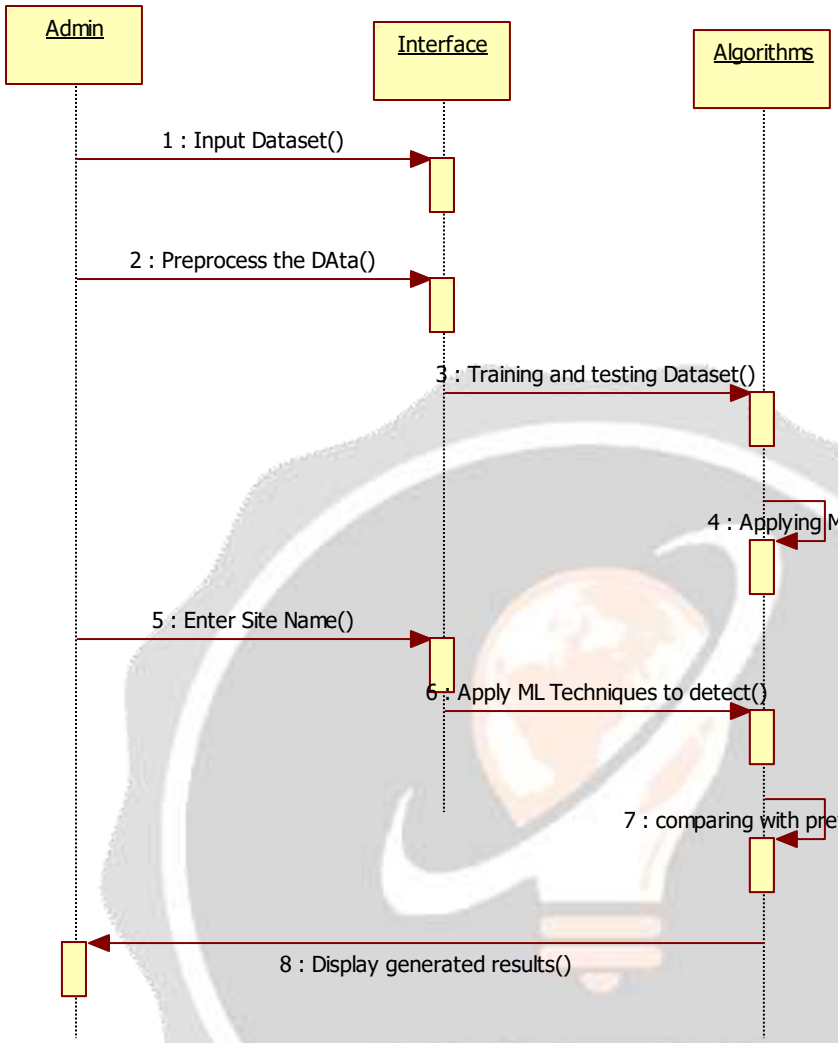
The Output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team members, in other words, the administrator and the clients. The output of VPN is the system which allows the project leader to manage his clients in terms of creating new clients and assigning new projects to them, maintaining a record of the project validity and providing folder level access to each client on the user side depending on the projects allotted to him. After completion of a project, a new project may be assigned to the client. User authentication procedures are maintained at the initial stages itself. A new user may be created by the administrator himself or a user can himself register as a new user but the task of assigning projects and validating a new user rests with the administrator only.The application starts running when it is executed for the first time. The server has to be started and then the internet explorer in used as the browser. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems can act as the clients. The developed system is highly user friendly and can be easily understood by anyone using it even for the first time.

3. LITERATURE REVIEW

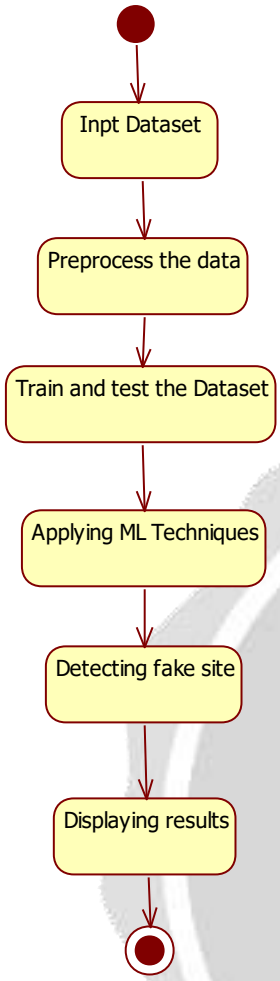
Phishing is a Web-based attack that seduces end users to visit fake websites and give away personal information such as user id and password. Phishing web pages are formed by fraudulent people to copy a web page from an original one. These phishing web pages are very similar to the original ones. Technical tricks and social engineering are extensively joined together for beginning a phishing attack. An important view of online security is to protect users from phishing attacks and fake website. Intelligent methods can be used to develop fake web pages. For this reason, internet users whether have enough experience in information security or not might be cheated. Phishing attacks can be launched via sending an e-mail that seems to be sent from a trusted public or private organization to users by attackers. Attackers get the users to update or verification their information by clicking a link within the e-mail. Other methods such as file sharing, blogs, and forums can be used by attackers for phishing. There are many ways to fight phishing including legal solutions, education, and echnical solution. Nowadays, information and communication tools are used in a manner that is very dense with information. For this purpose,various solution

3.1.PIPELINE FLOW

An important view of online security is to protect users from phishing attacks and fake website. Intelligent methods can be used to develop fake web pages. For this reason, internet users whether have enough experience in information security or not might be cheated. Phishing attacks can be launched via sending an e-mail that seems to be sent from a trusted public or private organization to users by attackers. Attackers get the users to update or verification their information by clicking a link within the e-mail.Other methods such as file sharing, blogs, and forums can be used by attackers for phishing. There are many ways to fight phishing including legal solutions, education, and technical solution. Nowadays, information and communication tools are used in a manner that is very dense with information. For this purpose, various solution methods for various problem types have been developed.In one form of interaction, a given use case may include another. "Include is a Directed Relationship between two use cases, implying that the behavior of the included use case is inserted into the behavior of the including use case.The first use case often depends on the outcome of the included use case. This is useful for extracting truly common behaviors from multiple use cases into a single description. The notation is a dashed arrowfrom the including to the included use case, with the label "«include»".There are no parameters or return values.



Architecture for Proposed Pipeline &Activity Diagram



Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.Activity diagrams are constructed from a limited repertoire of shapes, connected with arrows. The most important shape types:rounded rectangles represent activities;

- diamonds represent decisions;
- bars represent the start (split) or end (join) of concurrent activities;
- a black circle represents the start (initial state) of the workflow;
- An encircled black circle represents the end (final state).

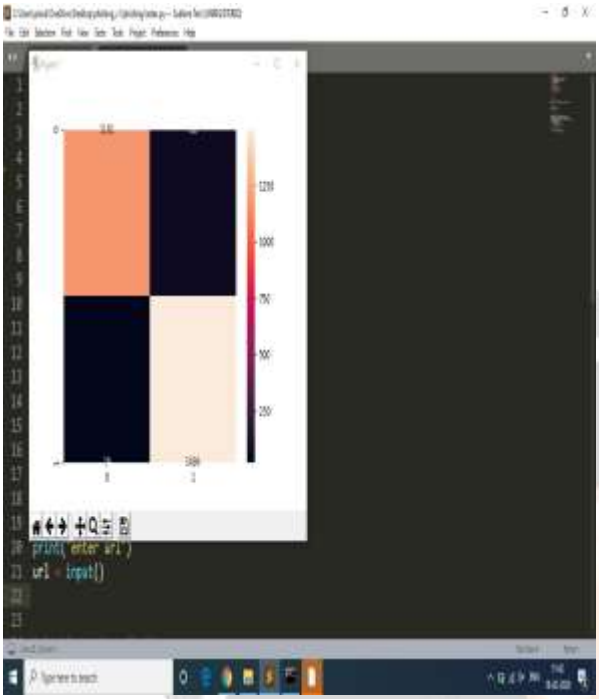
Arrows run from the start towards the end and represent the order in which activities happen. However, the join and split symbols in activity diagrams only resolve this for simple cases; the meaning of the model is not clear when they are arbitrarily combined with the decisions or loops

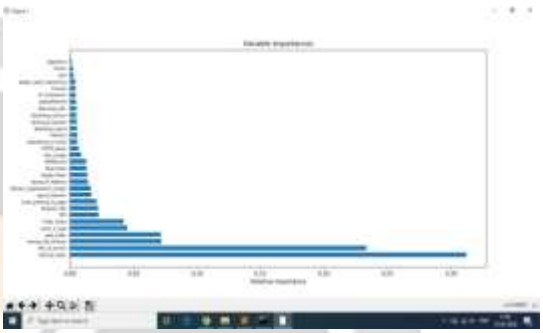
4. CONCLUSION AND FUTURE WORK

We defined features of phishing attack and thus proposed a model in order to classification of the phishing attacks. It consists of feature extraction from websites and classification section. In the feature extraction, we defined rules of phishing feature extraction and these rules have been used for obtaining features. Every user should also be trained not to blindly follow the links to websites where they have to enter their personal information. It is necessary to check the URL before entering the

website.

Sample output screenshot





Test Results confusion matrix

5. REFERENCES

[1]. N. Abdelhamid, A. Ayesh, F. Thabtah, "Phishing detection based associative classification data mining," Expert Systems with Applications, vol. 41(13), pp. 5948-5959, 2014.
[2] R. M. Mohammad, F. Thabtah, L. McCluskey, "Tutorial and critical analysis of phishing websites methods," Computer Science Review, vol. 17, pp. 1-24, 2015.
[3] H. Huang, S. Zhong, J. Tan, "Browser-side countermeasures for deceptive phishing attack," Fifth International Conference on Information Assurance and Security IAS'09, vol. 1, pp. 352-355, IEEE, 2009.
[4] R. M. Mohammad, F. Thabtah, L. McCluskey, "Predicting phishing websites based on self-structuring neural network," Neural Computing and Applications, vol. 25(2), pp. 443-458, 2014.
[5] M. A. U. H. Tahir, S. Asghar, A. Zafar, S. Gillani, "A Hybrid Model to Detect Phishing-Sites Using Supervised Learning Algorithms," International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1126-1133, IEEE, 2016.
[6] M. He, S.J. Horng, P. Fan, M.K. Khan, R.S. Run, J.L. Lai, R.J. Chen, Sutanto, "An Efficient Phishing Webpage Detector," Expert Systems with Applications, vol. 38(10), pp. 12018-12027, 2011.
[7] P. A. Barraclough, M. A. Hossain, M. A. Tahir, G. Sexton, N. Aslam, "Intelligent Phishing Detection and Protection Scheme for Online Transactions," Expert Systems with Applications, vol. 40(11), pp. 4697- 4706, 2013
[8] H. H. Nguyen, D. T. "Nguyen, Machine learning based phishing web sites detection," In AETA 2015: Recent Advances in Electrical Engineering and Related Sciences , pp. 123-131, Springer International Publishing, 2016.
[9] R. M. Mohammad, F. Thabtah, L. McCluskey, "Intelligent Rule-based Phishing Websites Classification, " IET Information Security, vol. 8(3), pp. 153-160, 2014.
[10] V. S. Lakshmi, M. S. Vijaya, "Efficient prediction of Phishing Websites Using Supervised Learning Algorithms," Procedia Engineering, vol. 30, pp. 798-805, 2012.
[11] J. James, L. Sandhya, C. Thomas, "Detection of Phishing URLs Using Machine Learning Techniques," International Conference on Control Communication and Computing (ICCC), pp. 304-309, IEEE, 2013.

[12] M. Al-diabat, "Detection and Prediction of Phishing Websites using Classification Mining Techniques", International Journal of Computer Applications, vol. 147(5), pp. 5-11, 2016.
[13] A. Hodzic, J. Kevric, A. Karadag, "Comparison of Machine Learning Techniques in Phishing Website Classification," 2016.