ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 4, April 2021

Classification Model for Phishing e-mails with a Datamining Approach

[1] Sharmila S P, [2] Balaji Rao Katika

[1] Assistant Professor, Dept. of Information Science and Engineering, Siddaganga Institute of Technology, Tumakuru
[2] Senior Research Scientist, V-Labs Private Limited, SLK group, Bengaluru
E-mail: [1] sharmila@sit.ac.in, [2] k.balaji@iitg.ac.in

***Abstract*----In the field of computer security, phishing attacks are one of the trending cyber-attacks. Phishing is an online criminal act that occurs when a malicious webpage impersonates as legitimate webpage so as to acquire sensitive information such as username, password, bank details by masquerading as a trustworthy entity. Attackers create a replica of an existing web page to fool users (e.g. e-mails, instant messages etc.). Phishing attack continues to pose a serious risk for web users and annoying threat within the field of electronic commerce. The increasing number of phishing websites has become a great challenge in e-business in general and in electronic banking also. The attacker makes a fake webpage by copying or making a little change in the legitimate page, so that an internet user will not able to differentiate between phishing and legitimate webpages. So, it is important to develop techniques which help in reducing these attacks. The theme of our project is to reduce the attacks by identifying them in the first place and avoid people to fall into such kind of traps.**

I. INTRODUCTION

In today's Internet world there is an enormous increase in the cyber-crime activities. Stealing one's identity is one among them. As per the Federal Trade Commission (FTC) details derived on 2020 [1], criminal affairs related to identity theft or stealing identity of a human is been hierarchically graded two with 20% of all customer grievances lodged. Phishing is process of usual way to steal an online user's personal identity, financial account credentials and other related data. This is also considered as "a fraudulent attempt, usually carried out through email, to steal or snip out individuals' personal or private information" [2]. It is a sort of crime which is employing technicalities along with social engineering. Many Social engineering schemes are consuming incautious or ungaurded victims as a prey by misleading and bluffing them with a certainity of belief that they are communicating with a highly trusted, legally rightful party, but through a deceitful email messages, email addresses or a website.

According to Anti-Phishing Working Group APWG [3] the over-all number of unique phishing Web sites detected as on September 2020 is 199,133. APWG has two major fundamental sources for phishing data: one among them is phishing emails conveyed to it by members of APWG and by members of the public, and the second one is phishing URLs conveyed by members of APWG into the eCrime eXchange of APWG. The APWG traces Unique phishing sites, Unique phishing e-mails focusses and also counts the number of brands attacked and confronted by investigating the phishing reports submitted.

The figure below shows the increase in the number of phishing sites from Aug 2019 to September 2020

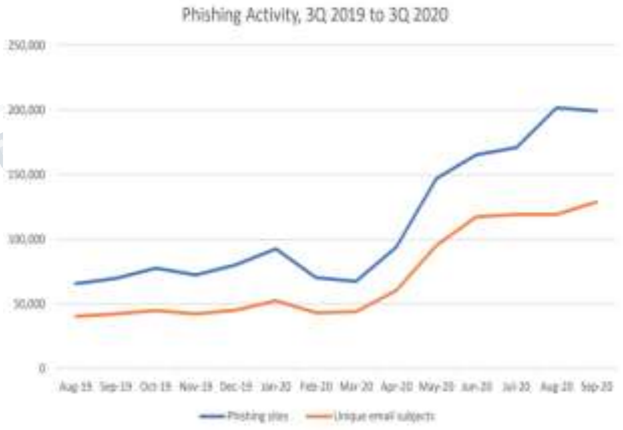

Fig.1: Statistics of Phishing Activity

When a new or fresh phishing email is not properly trained on the legacy sample data or blacklisted email it is referred as a zero phishing email attack. Existing phishing email detection schemes do not perform well against these attacks [4].

Different techniques adopted by phishers to initiate these phishing attacks are [5] email, smishing, instant messaging, voice phishing, Social network/media, malicious websites, search engines and spear phishing [8,9]

Most of the studies have shown that 65% of phishing attacks use email as the foremost means of communication channel to trap online customers [6,7]. We are focussing on building a detection method to detect phishing attacks through emails. We assume that, this is one of the greatest prevalent method used to ignite these attacks. Nonetheless, the text used in phishing emails can also be directed via some other channels of communication in order to reach online user. Likewise, the proposed and anticipated model can also be used to detect phishing emails at any other channels of communication providing the format used is HTML or text. Thus, this model is a momentous research contribution to determine and realize phishing attacks.

The main objective is to implement an algorithm to classify the emails, by extracting most significant features from the email content after pre-processing of emails and to train the model by applying extracted significant features from the existing dataset so that model could acquire the capability to classify if a new incoming email is ham or phish.

II. RELATED WORK

In one of the study analyses issued by APWG, there were at least 138,328 phishing attacks in the last 3 months of 2020. With a good proportion of research that has been carried out on anti-phishing in conniving numerous anti-phishing approaches. Afroz and Greenstadt (2009), characterised the current scheme of detection of phishing into three main types [9]: (1) non-content-based methods for classifying authentic or phishing mails without using site content, (2) content-based methods for catching a phishing mail using site contents and (3) similarity-based visual methods that consumes visual similarity with familiar sites to identify phishing.

AntiPhish is used to avoid users from using fraudulent web sites which in turn may lead to phishing attack. Here, AntiPhish traces the sensitive information to be filled by the user and alerts the user whenever he/she is attempting to share his/her information to an untrusted web site. The much effective elucidation for this is cultivating the users to approach only for trusted websites.

Table 1. Comparison of the Classifier Model-Based approaches

Authors	Contribution Summary	Weakness	Methodology
1)Chandra Sekran[10]	Structural features	Time consuming.	Support vector Mechanism
2)Ganger [11]	Training smart screen.	working with.fixed number.of features.	Bayesian statistics
3)Bazargani [12]	Uses heuristic way for text classification of phishing email	Levels.of accuracy.is low compared with.other techniques.	Semantic ontology concepts.
4)Chandra [13]	Phoney: Mimicking user response.	Collected data.are.so small in size. Time consuming.	PHONEY technique
5)Fette sadeh [14]	Pilferas prototypes.	Sizeable number of phishing and ham emails.	Random forest and support vector mechanism (SVM as a classifier).
6)Bergh Olz [15]	Statistical study.of phishing email	High memory requirement	Dynamic Markov chain
7)L.Mao Foghi [16]	Robust classifier model.	Using few numbers of Features.	Information gain algorithms.

III. ANTI-PHISHING

Countermeasures for phishing attacks are necessary to aim at preventing or detecting such attacks before or after victim data has been collected. Countermeasures for phishing detection and prevention separately can be categorized into five major categories − Machine Learning, Text Mining, Human Users, Profile Matching, and Others. The other category is further broken down into Ontology, Honeypot, Search Engine, and Client Server-based Authentication. [17]

Table 2. Some Anti phishing Techniques [18]

1.	One time password	Generated by various methods, it is valid for few minutes and can be used only once
2.	CAPTCHA	CAPTCHA: Completely Automated Public Turing test to tell Computers and Humans Apart. This method requires the genuine user to enter some information that is taken from a scrambled image This is usually difficult for automated robots to recognize and process.
3.	Hypertext transfer protocol secure	On the internet https is for secure communication. It is the result of simply layering the Hypertext Transfer Protocol (HTTP) on top of the TLS protocol, thus adding the security capabilities of SSL/TLS to standard HTTP communications.
4.	Digital Certificates	Digital certificates are used to authenticate users on both sides of communication. Many useful applications available in the banking domain[3]. This kind of authentication depends on the existence of a Public Key Infrastructure (PKI) and a Certificate Authority (CA), which represents a trusted third-party who signs the certificates attesting their validity.
5.	Attribute based anti-phishing techniques	implements both reactive and proactive anti-phishing defences.
6.	Genetic Algorithm Based Anti-Phishing Techniques	used to differentiate normal website from anomalous website by applying genetic algorithms. These anomalous websites refer to events with probability of phishing attacks.
7.	An Identity Based Anti-Phishing Techniques	This technique integrates partial credentials sharing and client filtering techniques. It prevents phishers from easily masquerading as legitimate online entities.
8.	Character Based Anti-Phishing Approach	This technique uses characteristics of hyperlink in order to detect phishing links. Linkguard is a tool implements this technique. After analysing many phishing websites, the hyperlinks can be classified into various categories with this tool
9.	Content Based Anti-Phishing Approach	This mechanism gives higher rank to well-established web sites. It has been observed that phishing web pages are active only for short period of time and therefore, will acquire low rank during internet search and this becomes basis for content based anti-phishing approach.

IV. BACKGROUND CONCEPTS:

Support vector machine (SVM): It is one of the supervised learning algorithms. It belong to both the regression and classification categories of machine learning algorithms by using associated learning algorithms.. It can be used to recognize patterns in given data or given a set of training data it is intended to classify. separating data into training and testing sets is the major task involved in A classification model. SVM produces a model (based on the training data) which would predict the target values of the test data. This method does not suffer the limitations of data dimensionality and limited samples. Several recent studies have reported that the SVM generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. It has been employed in a wide range of real-world problems such as text categorization, digit recognition, hand-written, image classification, tone recognition, micro-array gene expression data analysis, data classification and object detection. Finally, it can be concluded that SVM acts as a machine learning based system for the detection of malware and phishing sites [2].

Bayesian statistical methods use Bayes' theorem to calculate and apprise probabilities after obtaining fresh data. Bayes' theorem can be used for the estimation of the parameters of a probability distribution or statistical model. Bayesian statistics just treats probability as a degree of certainty or acceptance, Bayes' theorem directly assigns a probability distribution that measures or quantifies the certainty to the parameter or a set of parameters. Let P(x) be the prior probability of A which expresses one's certainty about occurrence of event A before evidence is taken into account. The prior probability may also quantify prior knowledge or information about A. P(A|B) is the likelihood.

$P(B) = P(B|A_1)\ P(A_1) + P(B|A_2)\ P(A_2) +\ldots + P(B|A_n)\ P(A_n)$

$= \sum(k=0)^n P(B|A_n) P(A_n)$.

Semantic Ontology: Conventional and conservative detection techniques find difficulty to keep up the performance with the intrinsic complexity of web application design and hence there is a ever-growing variety of attacks that can always exploit it. There is a need of Security frameworks that are modelled with an ontological approach in order to promise new stripe of defence that can be highly operative and effective in detecting zero-day phishing and sophisticated web application attacks. Such security frameworks can also capture the context of the actual contents of information such as in-line scripts, HTML pages and have the ability to sieve these contents by taking into contemplation their consequences to the target applications. We can say that an ontology-engineering methodology can be systematically functional for designing, estimating and evaluating security framework systems.

Random forest: Random forests are an ensemble of or collection of learning methods for classification, regression and other tasks that activates and operates by constructing an assembly or multitude of decision trees at training duration. Then output the class that is the mode of the classes by means of classification or mean prediction by means of regression of the individual trees. Random decision forests always meant for correcting the decision trees' habit of always getting overfitting to their training set.

Dataset:

Dataset considered is from spam assassin website [3]. It comprises of 5 sets of emails datasets. 2 set of easy hams, 1 set of hard hams and 2 sets of spams.

V. SYSTEM ARCHITECTURE

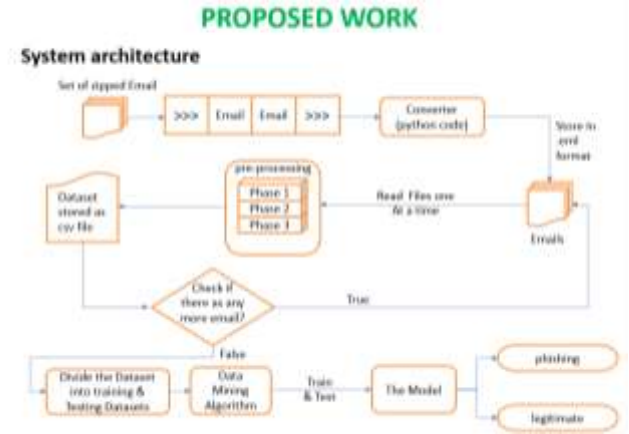

Fig 2. System architecture

The Figure 2 depicts the complete architecture of the proposed system. The process of converting the emails to .eml format is achieved through the converter module. During the preprocessing emails are read one at a time then the dataset is stored as a .csv file. This process is carried out till no emails are left in the dataset. The actual dataset is segregated into two groups for training and for testing before the application of the Data mining algorithm. Finally, the classification model successfuly classifies the legitimate mails and phishing mails.

VI. VI DESIGN OF PREPROCESSING PHASE OF THE PROPOSED SYSTEM:

Implementation of the work is done in three phases as follows:

Phase1: Init Pre-processing
Phase2: Feature Extraction
Phase3: Training of System using Machine Learning Algorithm

Phase 1: Init **Pre-processing**

In this phase undesirous elements in the emails that is not mandatory for feature extraction are detached and they are also converted to a proper format in this phase. Error prone situations are common if the text file is processed for feature extraction, this would also hinder the overall performance.

Pre-processing phase is carried out in the following steps:

Since the Dataset downloaded will be in a unsupported format it needs a convertion before processing. converting the files into .eml format is found to be effective. So that header, body and html part can be separated easily from each email sample. URLs and the text along with it can also be extracted and kept aside for each emails in the dataset. Href and URLs from html part of email is extracted and retained in a separate file.

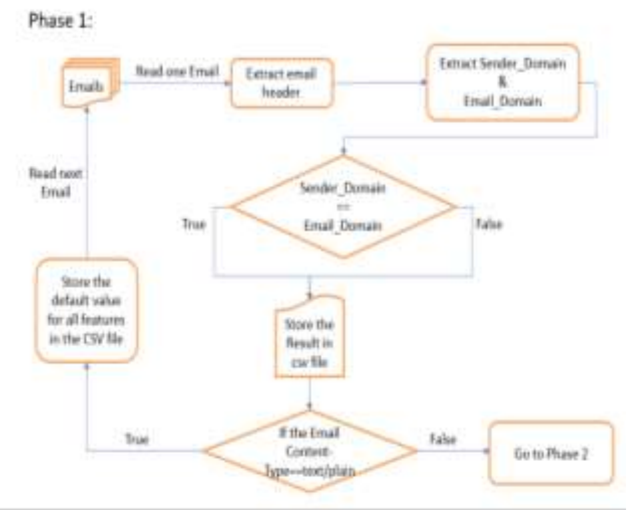
Phase 1:
Emails
Read one Email
Extract email header
Extract Sender_Domain & Email_Domain
Read next Email
Sender_Domain == Email_Domain
True
False
Store the default value for all features in the CSV file
Store the Result in csv file
True
If the Email Content-Type==text/plain
False
Go to Phase 2

Fig 4. Phase 1: Init Preprocessing

Phase 2: Feature Extraction

After preprocessing the next phase is feature extraction. By extracting the body of the email there is a need to search for the following features in the body. 1)HTML form in the email 2) java script in the email 3) disparities between href and link text 4) Number of dots in the domain name 5) number of links present or embedded in the email 6) IP based URLs in emails 7) word list features pertaining to frequency of usage

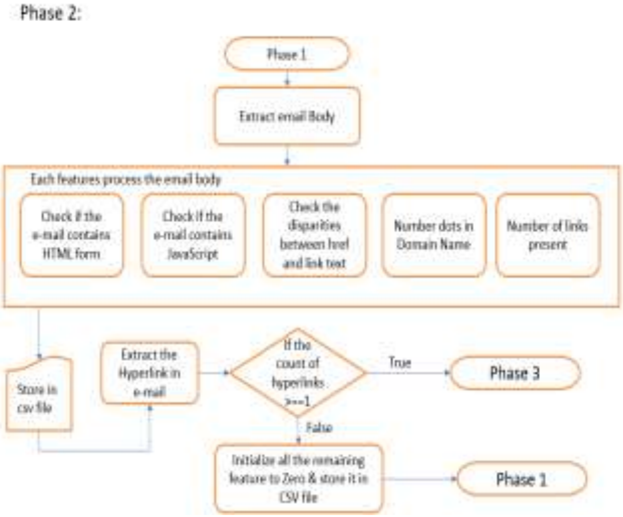
Phase 2:
Phase 1
Extract email Body
Each features process the email body
Check if the e-mail contains HTML form
Check if the e-mail contains JavaScript
Check the disparities between href and link text
Number dots in Domain Name
Number of links present
Store in csv file
Extract the Hyperlink in e-mail
If the count of hyperlinks >=1
True
Phase 3
False
Initialize all the remaining feature to Zero & store it in CSV file
Phase 1

Fig 3. Phase 2

Part of the E-mail message
Header Part
Components of Email Message
Two parts Of Email
Header part
Body part
Body part or Contents Part
@

Fig 5. Components of Email

Phase 3: Training of System using Machine Learning Algorithm

After preprocessing and feature extraction the next stage is to train the system for classifying the given email is a phishing email or not which is the core objective of the work. Random Forest is used to achieve the objective. Random forest (RF) is the famous learning classification and regression method suitable for experimenting on problems which usually involve grouping of data into various classes. In RF, prediction and estimation is achieved with decision trees. During the training phase, a number of decision trees are constructed which are then used for the class prediction. This is achieved by considering the votes or nominated classes of all the individual trees and the class with the highest nomination or votes is only considered as the required output.

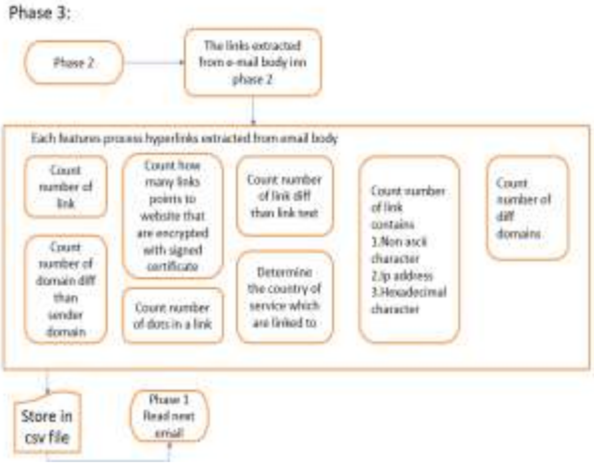
Phase 3:
Phase 2
The links extracted from e-mail body inn phase 2
Each features process hyperlinks extracted from email body
Count number of link
Count how many links points to website that are encrypted with signed certificate
Count number of link diff than link text
Count number of link contains 1.Non ascii character 2.Ip address 3.Hexadecimal character
Count number of diff domains
Count number of domain diff than sender domain
Count number of dots in a link
Determine the country of service which are linked to
Store in csv file
Phase 1 Read next email

Fig 6. Phase 3

VII. RESULTS

Phishing a very serious threat to world wide or global security and economy which cannot be ignored and need to be addressed for resolution immediately. With the advent of new phishing websites in a fast rate and emergence of numerous distributed phishing attacks has caused it tough to update the blacklists database. Therefore, we are making an attempt by presenting a phishing detection approach which is content-based, this will bridge the current gap that is acknowledged in our literature survey.

After rigorous review of the necessities, need and the results conforming to the testing dataset, it has been established and recognised that our application can successfully categorise emails with the best accuracy of 96.23%, whereas the inclusive accuracy lies between 95% and 96% in the entire testing dataset.

Algorithm: To classify Phishing Emails

```
Begin_Phish_Algorithm
Input:
N be the number of nodes of tree
M be the number of features
D be the number of trees to be constructed
Output:
V be the class with the highest nominations or votes
while stopping condition is false
do
Randomly or haphazardly consume a bootstrap sample
A from the training data D
Use the following steps given below to construct a tree
Ti from the consumed bootstrapped sample A:
(I)   select Randomly m features from the set M; where
      m≪M
(II)  For node d, compute the finest split point amongst the
      m features
(III) Split the node into two child nodes using the best split
(IV)  Repeat I, II and III until n number of nodes has been
      gotten
(V)   Similarly Build your entire forest by repeating steps
      from I–IV for D number of times
end while
output all the trees constructed {Ti} 1D
apply another new sample to each of the constructed
and existing trees starting from the root node
assign the sample to the class matching to the leaf
node.
combine and associate the decisions (or votes) of all
the trees Output V, that is, the class with the highest
vote.
end_Phish_Algorithm
```

VIII. CONCLUSION AND FUTURE WORK

By looking into the performance and accuracy we can conclude that, though the accuracy seems to be pretty good, for real world examples, it might fall down to classify the new incoming email correctly, due to false negative and false positive constraints. This downside flaw can be pragmatically observed and experimented in a full-fledged and sophisticated implementations of the current perception on Outlook, Gmail, etc. It is the universal truth that No product can be considered as the definitive. There is always a scope for perfection. Similarly, this system can be greatly enhanced by the accumulation of a more significant features for extraction and training it with more diversified and assorted dataset. The proposed system can be fine-tuned before embedded into mail server so that it can be applicable for factual time classification model that can be deploying onto email clients. The overall performance of the proposed algorithm can still be amended and enhanced through healthier training..

REFERENCES

[1] Federal Trade Commission, et al. Consumer Sentinel Network Data Book for January–December 2015, 2016.

[2] L. OpenDNS, PhishTank, 2016, https://www.phishtank.com/index.php.

[3] APWG, Phishing activity trends report, 3rd Quarters 2020, Report, APWG. Activity July-September 2020 Published 24, November 2020. https://docs.apwg.org/reports/apwg_trends_report_q3_2020.pdf

[4] A. Almomani, T.-C. Wan, A. Manasrah, A. Altaher, M. Baklizi, S. Ramadass, "An enhanced online phishing e-mail detection framework based on evolving connectionist system", International Journal of Innovative Computing, Information and Control (IJICIC) 9 (2013) 169–175.

[5] Sami Smadi, Nauman Aslam, Li Zhang "Detection of online phishing email using dynamic evolving neural network based on reinforcement learning" Department of Computer and Information Science, Northumbria University, UK

6] A. Vishwanath, T. Herath, R. Chen, J. Wang, H.R. Rao, "Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model", Decision Support Systems 51 (2011)576–586.

[7] C.L. Tan, K.L. Chiew, K. Wong, "PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder", Decision Support Systems 88 (2016) 18–27.

[8] K. Krombholz, H. Hobel, M. Huber, E. Weippl, "Advanced social engineering attacks", Journal of Information Security and Applications 22 (2015) 113–122.

[9] D.D. Caputo, S.L. Pfleeger, J.D. Freeman, M.E. Johnson, "Going spear phishing: exploring embedded

training and awareness", IEEE Security & Privacy 12 (2014) 28–38.

[9] Afroz and Greenstadt (2006). "A machine learning Approach to phishing Detection and defence".

[10] M. Chandrasekaran, et al., "Phishing email detection based on structural properties", in New York State Cyber Security Conference (NYS) , Albany, NY ," 2006

[11] P. R. a. D. L. Ganger, "Gone phishing: Evaluating anti-phishing tools for windows. Technical report," September 2006.

[12] M. Bazarganigilani, "Phishing E-Mail Detection Using Ontology Concept and Nave Bayes Algorithm," International Journal of Research and Reviews in Computer Science, vol. 2, no.2, 2011.

[13] Elsevier B.V(2015) "An ideal approach for detection and prevention of phishing attack", 4th international conference on advances in computing, communication and control.

[14] M. Chandrasekaran, et al. (2006), " Phishing email detection based on structural properties", in New York State Cyber Security Conference (NYS), Albany, NY,".

[15] A. Bergholt, et al., "Improved phishing detection using model-based features," in Proc. Conference on Email and Anti-Spam (CEAS). Mountain View Conf, CA, aug 2008.

[16] L. Ma, et al., "Detecting phishing emails using hybrid features," IEEE Conf, 2009, pp. 493-497.

[17] Ahmed Aleroud, Lina Zhou, "Phishing environments, techniques, and countermeasures: a survey"

[18] L. Joy Singh, "A Survey on Phishing and Anti-Phishing Techniques", NIELIT IMPHAL IJCST Mar-Apr 2018