International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 07 | July -2020 ISSN: 2582-3930

Analysis different algorithms used for detection of email phishing

Pooja Suresh Kadam[1], Prof Mansi Kambli[2]

[1]*PG M.Tech Student, Computer Engineering & K.J.Somaiya College of Engineering, Mumbai.*
[2]*Professor ,Computer Engineering & K.J.Somaiya College of Engineering, Mumbai.*

Abstract - Phishing attacks are widely spread and most common online now a days. This attack is growing at an alarming rate these days. These attacks have resulted in financial losses to many individuals, companies etc by social engineering. Even though this attack is so common phishing email detection is not only needed but it has to be detected with great accuracy. Many machine learning algorithms are proposed to detect phishing emails but accuracy of those algorithm is not good. So in this paper different classifiers that are used for phishing emails detection are analyzed and their results are compared. Different classifier used here are Bayesian classifier, Support Vector Machine (SVM) and Cuckoo search-Support vector machine (CS-SVM). Here Bayesian classifier and SVM is content based classifier, and CS-SVM is hybrid classifier. Based on the accuracy, precision and error best classifier is chosen.

***Key Words*:** Phishing attack, Email, bayesian classifier, SVM, CS-SVM, detection.

1.INTRODUCTION

The multiplied use of social media technologies has created communication easier and quicker. However there are several attacks that are performed through internet to achieve sensitive information or to hurt the image of an individual or company. One amongst dangerous attack is the attack is "Social engineering attack" .The art of influencing individuals to reveal sensitive information is called engineering attack and the process of doing so is known as social engineering attack [2]. They first analyze the weakness and then try to exploit it, victim can be single person of group of people. Using phone calls and other media, these attackers trick people into handing over access to the organization's sensitive information. There are different types of social engineering attacks, one of them is phishing attack.

In phishing attack an individual or group of people (organization) called as phishers try to steal personal information from victim such as passwords, bank account number, personal identification number etc. Phishing attacks are of different types and can be done through different medium. Various types of phishing are vishing, smishing, search engine phishing, spear phishing etc. Here in this paper methods used for email phishing detection are analyzed, many studies have been proposed and algorithms are implemented to detect phishing attacks. There are different approaches used to detect phishing but here the best approach with high accuracy is found out by analyzing 3 different algorithms.

In this paper comparison of different machine learning algorithms is done and results are analyzed. Three algorithms compared here are Naïve Bayesian [3], SVM [5] and CS-SVM [4].

2. LITERATURE SURVEY

Phishing attack is a cyber crime where the attacker tries manipulates an individual or group of people to share their personal information or data, it is huge and serious security issue in the society. These attack is generally performed by emails, also termed as spear phishing attack. In this attacker the attacker makes harder for victim to distinguish between legitimate and phishing or spam email. There are various types of phishing attacks like Spoofing email, Fake Social Network Accounts, Hacking, Trojan horse etc. In [1] different types of phishing attack are mentioned and solutions used for prevention of attacks is also discussed. Apart from this prevention methods discussed there are many machine learning algorithms which are used in detection of phishing attack. The 'art' of influencing people to divulge sensitive information is known as social engineering and the process of doing so is known as a social engineering attack. The protection of information is extremely vital situation in this modern society. Even though the reliability and security around information is continuously improving, a liability that still remains is the human actors who are vulnerable to manipulation techniques. In [2] social engineering as a cyber crime domain and detection techniques of social engineering attack as a process inside this is explained in domain. The Social Engineering Attack Detection Model(SEADM) model provided the common

procedural arrangement for implementing detection mechanisms for social engineering attacks. The state diagram provides an additional abstract and extensible model that highlights the inter-connections between task categories related to totally different scenarios. One of the algorithm used in comparison is Bayesian classifier [3]. It is used here to classify if the email is phishing or legitimate using supervised learning across feature extraction. Here Content based filter checks if there is any text within the body of Email, then URL and additionally it also considers the mail header as a feature, this mail header's subject part is used for classification of text . Text classification task is performed by preprocessing the TEXT with various terms and conditions. Those terms can be Stopword Removal, Word frequency calculation to determine word probability, tokenizing and HTML tag removal determines if mail is spam or not etc. Hybrid classifier [4] is used here to detect if the mail is phishing or legitimate. One of the algorithm used is Support Vector Machine (SVM) and other is Cuckoo search(CS) algorithm. 23 features from the body of email, header of email and URL present in the email are extracted. Then the Cuckoo Search (CS) algorithm is basically used for parameter selection of kernel operations. The hybrid classifier in [4] combining Cuckoo Search(CS) algorithm with SVM is evaluated on a training and testing dataset including both old and new phishing emails and yields a higher probability of obtaining better results than just SVM classifier. This is the best classifier with highest accuracy and lowest false positive rate. Hybrid features [4] are used here that are content based, behavior based and URL based. First run a group of scripts that extracts all the proposed features from each and every incoming email in the form a feature value that is o or 1. After the features are extracted, those value are used to train the classifier and classify whether the email into phishing email and legitimate email. Here SVM classifier [4] is used as it is the most popular technology of all the content-based approaches. Machine learning-based technique has been shown to be effective to detect phishing email.

3. CHALLENGES IN EMAIL

There are security issues present in various part of email that help attacker to attack target without knowing or any clue. There are different parts of emails like email header, email body and url present in the email must be checked and feature must be extracted.

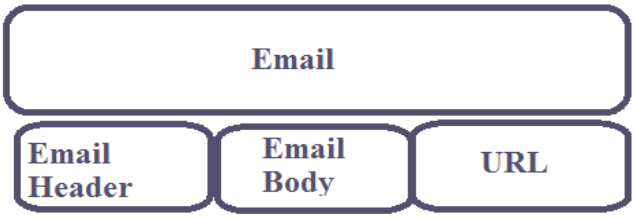

Fig -1. Features to check if Email is phishing

1. Email header based security issues: Here email header based features are checked.
   Number of CCs, Presence of dark copy, whether the sending time of email is normal work time or not, if the mail is a reply email or not, presence of suspicious words in subject like order, payment, re etc.S
2. Email body based security issues: Here email header based features are checked.
   If sensitive words are present in the LINK text, presence of javascript, variance between LINK text and href attribute.
3. URL based security issues: If the email contains any url present in it then that url must also be checked.
   If url consists of IP address, total number of dots in present in domain name is greater than 3, presence of "@" symbol in urls, If length of url is greater than 54 characters then url is legitimate, total number of http/https present in an url must not be greater than 1.

4. COMPARISON OF ALGORITHMS

4.1 Bayesian Classifier

In [3] Bayesian classifier is used for content based phishing email detection. As Bayesian classifier is popular statistical classifier, it uses text classification method for identifying spam mails. In text classification bag of words feature is used. Bayesian used tokens for spam and ham mails to calculate probability and determine whether a mail is phishing or not. Bayesian steps and equation for phishing email detection:

1. Class prior probability : P (phishing)
2. Likelihood of email: P (word/phishing)
3. Posterior probability: P (phishing/word)
4. Predictor prior probability : P (word)

Final equation will be:

P (phishing/word) = [P (word/phishing) P (phishing)] / p(word)

Architecture Bayesian for phishing email detection :

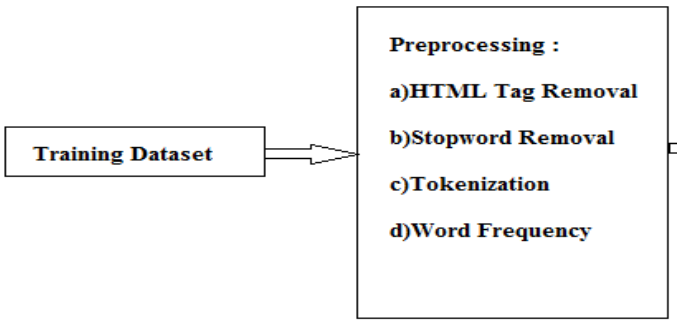

Fig -2. Architecture of Bayesian classifier [3]

Training dataset is collection from Gmail it consists of both phishing as well as legitimate mails. These mails are considered to be the input in HTML format for preprocessing. In preprocessing the unwanted noise in the mail is removed. This removed data any how does not help in calculating accurate results.

There are 4 types in it:

1. Removal of HTML Tag: As input or loaded mails are generally in HTML format it holds tags that are not used in classification, so this text need to be purified by removing all tag.
2. Removal of Stopword: Stopword are the list of words which contains certain high frequency words, conjunctions, prepositions and terms including articles.
3. Tokenization: Tokenising is also known as lexical analysis, here content in the text is divided into different strings of characters also known as Tokens. In filtering, techniques like white or blank space removal and punctuation symbols are removal is done in tokenization.
4. Word Frequency: In word frequency, frequency of words is calculated depending on its occurrence, this helps in deriving probability of the word for being phishing or legitimate mails.

After preprocessing is done Bayesian classifier is method of text classification. So the algorithm evaluates both phishing as well as legitimate mails and then gives performance measurement on the basis of classification. Next is evaluation of test dataset, here after training of Bayesian classification is done testing dataset is preprocessed and classified using trained classifier. Last is Performance Evaluation, Here all the positive and negative values are evaluated. And from obtained values accuracy is calculated.

4.2. Support Vector Machine

Support vector machine (SVM) is supervised learning algorithm, it is set of related methods used in classification. SVM algorithm is also known as maximum margin classifier as it can parallel perform tasks like minimize the observed classification error and maximize the geometric margin.

In [5] hybrid features are used with SVM classification to detect phishing email. Here different features like domain name of sender, blacklisted words in subject and content, total number of dots in domain name is must be greater than 3, IP address in URL, whether a mail is reply email, etc features are checked. All these features are explained in 3.2 Security Issues.

Architecture of SVM for phishing email detection:

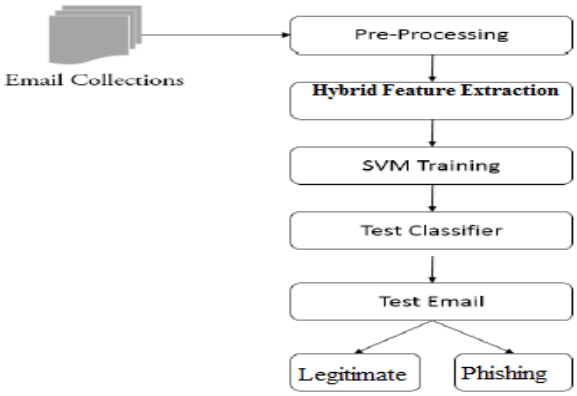

Fig -3. Workflow architecture of SVM [5]

Email collection is dataset of emails that are to be classified if phishing or legitimate.
First is pre-processing is used here to remove unwanted terms from email that are irrelevant for classification and do not help in giving accurate results. This step includes detection and elimination of numbers, ,stripping HTML, special symbol like @, word stemming. Secondly hybrid feature extraction is used for extraction of important and relevant features for classification of email from the email body. Next is SVM training, here phishing email training dataset is used to train classifier. After the classifier is trained it is used as test classifier to test and classify the data in test dataset. Last is testing of single email after the training and testing phases are completed, single email is given as input to the classifier and based on training of dataset classifier classifies the email. The output generated is either in the forms of 0 or 1, 1 represents that mail is phishing and 0 represents mail is legitimate or not a phishing.

4.3. Cuckoo search-Support Vector Machine

Cuckoo search-SVM (CS-SVM) is a hybrid classifier consisting of cuckoo search and svm. In [4] as svm algorithm works by parameter selection in kernel function, here cuckoo search algorithm is integrated with svm as it optimizes the parameter selection in kernel function and gives better results. CS algorithm uses Levy flights for generation of step size and effectively search the solution space.

Here in CS-SVM Cuckoo search is used for optimizing parameter and svm algorithm is used as fitness function, here value of CS is used to generate the hyperplane that minimizes error and maximizes margin with correctly classified data points. With the help of hyperplane that is constructed these error are classified as either legitimate or phishing emails. CS algorithm is modified and changes are executed either until classification error is unchanged or until iteration reaches its maximum number of limits.

Architecture of CS-SVM:

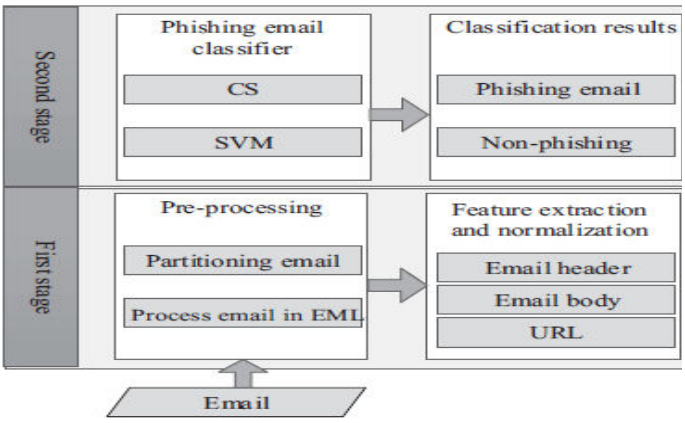

Figure.4. Architecture of CS-SVM [4]

In figure. pre-processing phase where conversion of mail into XML format takes place and email is split into three parts that are head, body and url of email. Step 2 is feature extraction, here various features are extracted from email head, email body, email url and feature standardization. There are different features extracted from different part of email. All these features are mentioned in III. Step 3 is use of CS-SVM classifier for classification of phishing emails where CS is used for optimizing parameters of kernel function in svm. Step 4 is classification result here the classification is completed and phishing emails and legitimate emails are separated.

5. COMPARATIVE STUDY

In this section comparison between different classifiers is done. All three classifiers compared here are explained in above section. Comparison is done based on parameters like accuracy, precision, recall and error.

Table -1: Comparison between different algorithms

Sr.no	Algorithm	Feature approach	Accuracy	Precision	Recall	Error
1.	Bayesian classifier	Content	93.98%	0.93%	0.95%	6.02%
2.	Support Vector Machine(SVM)	Content	97.75%	95.65%	99%	2.75%
3.	Cuckoo search-Support Vector Machine(CS-SVM)	Hybrid	99.52%	100%	93%	0.79%

From Table.4 we see that Bayesian classifier and SVM are content based approach while CS-SVM is hybrid approach. In terms of accuracy and precision CS-SVM is higher than other with 99.52% and 100% respectively. But the recall of SVM is higher. In error column the classifier containing least error is CS-SVM with 0.79% of error rate.

So from all the comparison made, we observe that CS-SVM is the best of the three classifier.

6. CONCLUSIONS

So in this paper comparison of three different classifier is done having different feature approaches such as content classifier and hybrid classifier. Classifiers compared here are Bayesian classifier, SVM and CS-SVM. Based on the analysis made by comparison made between three algorithms the accuracy and precision of CS-SVM is highest with 99.52% and 100%. All the features that were to be checked as security issues are checked in this classifier. This hybrid approach uses cuckoo search algorithm with SVM classifier to select optimal parameter value in kernel function evaluates all the security issues that are email header based, email body based and URL based. It can be concluded that hybrid approach using classifier that is CS-SVM has highest accuracy, highest precision and lowest error. Therefore this classifier is the best among three classifiers

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my guide “Prof. Mansi Kambli” sir for their able guidance and support in completing my project.

I would also like to extend my gratitude to the Principal madam “Dr. Shubha Pandit” and HOD of Computer Department Sir “Dr. Deepak Sharma” for providing me with all the facility that was required.

REFERENCES

[1] Gupta, Surbhi, Abhishek Singhal, and Akanksha Kapoor. "A literature survey on social engineering attacks: Phishing attack." 2016 international conference on computing, communication and automation (ICCCA). IEEE, 2016.

[2] Underlying finite state machine for the social engineering attack detection model

[3] Rathod, Sunil B., and Tareek M. Pattewar. "Content based phishing detection in email using Bayesian classifier." 2015 International Conference on Communications and Signal Processing (ICCSP). IEEE, 2015.

[4] Niu, Weina, et al. "Phishing Emails Detection Using CS-SVM." 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC). IEEE, 2017.

[5] Form, Lew May, Kang Leng Chiew, and Wei King Tiong. "Phishing email detection technique by using hybrid features." 2015 9th International Conference on IT in Asia (CITA). IEEE, 2015.

[6] Mouton, Francois, et al. "Underlying finite state machine for the social engineering attack detection model." 2017 Information Security for South Africa (ISSA). IEEE, 2017.

[7] https://www.tripwire.com/state-of-security/security-awareness/5-social-engineering-attacks-to-watch-out-for/. [November.17, 2019]

[8] Shradhanjali, Prof. Toran Verma“E-Mail Spam Detection and Classification Using SVM and Feature Extraction”

[9] https://www.tripwire.com/state-of-security/security-awareness/5-social-engineering-attacks-to-watch-out-for/. [November.17, 2019]

[10] https://www.social-engineer.org/framework/attack-vectors/attack-cycle/ [November.15, 2019]

[11] https://artint.info/html/ArtInt_181.html [November. 16, 2019]

[12] https://www.sciencedirect.com/science/article/pii/S2210832717301679 [November. 16, 2019]