APPROACHES TO PHISHING ATTACK DETECTION USING NATURAL LANGUAGE PROCESSING: A SURVEY

[1]Ruchi S. Ghantkar, [2]Nilesh B. Fal Dessai

[1]PG Student, [2]Head of Department,
[1,2] Department of Information Technology,
[1,2] Goa College of Engineering, Farmagudi, Goa, India.

Abstract : Phishing attack is a cyber-attack carried by a fraudulent attacker who impersonates to be a trustworthy person to defraud of the money or personal information of the user such as login credentials, ATM card details through some electronic channels such as email, text messages, voice calls etc. by luring the user to click on the link in the email. With the rise in technology, phishing attacks are also rapidly increasing these days. There are number of approaches that have been proposed to protect the systems from such unwanted phishing attacks. The aim of this paper is to survey various approaches for detection of phishing attack using natural language processing.

Index Terms **- Phishing attack, NLP, Natural language processing, SVM, Support vector machine, Email frauds**

I. INTRODUCTION TO PHISHING ATTACK

Phishing attack is a type of cyber-attack that is well planned and executed, by a fraudulent attacker with an intention of stealing money or sensitive information. Such attacks are done through the unexceptional medium channels like emails, SMS, voice calling etc. Phishing attack is a very common cyber attack which was first put to an end by an internet provider company called America Online in the year 1995 [13]. Same as in fishing activity, the fishes fall prey to the bait, the attack is named as phishing because a legitimate person falls prey to the bait of the attacker. The victims are lured by some unpromising schemes and eventually they fall prey to the attacker, thereby exposing the personal or corporate information to the attacker. The cyber attackers target the people, rather than breaking the hardly secured computer system. Targeting the humans is easier than strong computer systems for the cyber attackers. People are more vulnerable to such attacks because the phishing emails seem to be ordinary but they are fraudulent emails in nature. The main reason behind the lack of awareness is due to the lack of education or training. For example, the mail received looks to be legitimately sent by the employer to their employee related to the official task and hence login is mandatory, thereby the attacker deriving the password to further steal information of the company. All manner of security systems can be outwitted if the user divulges the password or any other critical information. Phishing emails, requesting for the sensitive information fraudulently, are a common way of the attack, unlike social engineering, which is formed to utilizes psychological weaknesses of the victim.

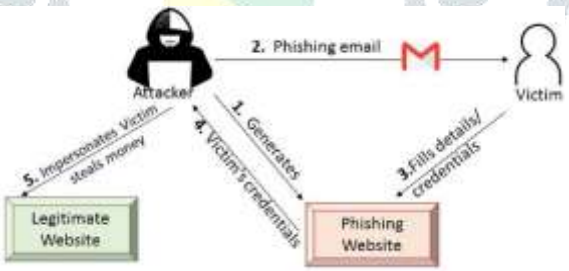

Fig 1 Phishing attack mechanism

The phishing attack processes in five steps as shown in Fig. 1

[1] Attacker generates a phishing website. The fraudulent website replicates the actual legitimate website to manipulate the user into trusting that the website is safe. The website has a form which is designed to get the user information like credit card details, user details, passwords etc.

[2] Attacker tries to find out which emails can possibly be targeted for the attack. Large number of emails are then sent to all the targeted victims with URL of the fraudulent website

[3] User receives the email. User clicks on the given link and fills the details. For example, the fake website displays a login section to fill user id, password and then click Login button.

[4] In the above example, once the user clicks the login button, this process internally sends the details to the attacker.

[5] Attacker impersonates and logs in to the real account of the victim, with this received details, the former transfers the money to his/her own account. Impersonation is the method of trying to be some legitimate entity. Impersonation of some genuine user or genuine website is done to gain the trust of the victim.

An increase of phishing attacks by 65% in the year 2016 was seen compared to year 2015 which accounts in more than 1.2 million, was recorded by The Anti-Phishing Working Group [11]. 1.4 million Phishing sites were created per month, as estimated by an organization named Webroot. The Federal Bureau of Investigation (FBI), research on Business E-Mail Compromise in 2018, studied that due to the phishing attacks on the emails the businesses had to cost $12 billion world wide.

1.1 TYPES OF CYBER ATTACKS

[1] Vishing: Voice Phishing or Phishing attack over Voice. The data is collected from Voice Calling. We have seen multiple cases where a fraudulent person calls and asks the users about their OTP, Passwords and personal information on the call, by faking about the organization they belong to. Example, making user believe the user that they are calling from a particular bank and gain user's trust.
[2] Smishing: SMS phishing, text SMS is the medium to collect information from the targeted victim.
[3] Evil Twins: Method used to collect information from the fake website. This is carried at public location like Airport, Railway station etc.
[4] Spear Phishing: This is a most common type of phishing method where a fake website is sent to the victim via email channel. Information is collected from other social media platforms and a target victim list is selected
[5] Whale Phishing: The type of phishing attack that attacks high targets like some big personality like board members of a company.
[6] Dropbox: Dropbox is an online storage service where company workers store the files. Phishing attackers create a fake Dropbox login, thereby confusing the employees while login.

1.2 SPEAR PHISHING

Spear Phishing is a very common type of cyberattack. As email is an unchallenging medium for executing the cyberattacks, it is widely used and hence the crime rates in this field are increasing rapidly. The fraudulent attacker selects a target victim or the list of target victims and sends the phishing email. To know whether the email is legitimate or phished, is crucial for the user. Hence having a detection system for the email is a must. As of now, various manners of phishing email detection systems have been developed. The methodology of detection is generated using various technologies like blacklist, whitelist, URL detection, header analysis, content analysis, detection using natural language processing etc. In this paper, we have studied some of the methodologies conducted to detect the phishing email using the natural language processing.

II. NATURAL LANGUAGE PROCESSING

A natural language is the language in which humans communicate via speech or text. Natural Language Processing is a technique of machine learning, where a machine can understand human language. Using NLP, the machine is able to perform text reading and understanding, speech recognition, speech translation and obtaining meaning from the human language.

A tremendous amount of data formed each day. Data is generated from communication platforms such as various social media platforms like Twitter, Facebook, WhatsApp, Instagram etc. Majority of the data exists in the textual form. About 21% of the same is a structured data and remaining is unstructured in nature, according to the industry estimates. The aim is to essentially process the text into textual data analysis via the application of natural language processing. Text Mining is the process of deriving the significant information from the text. Text mining plays a major role in the text analysis in natural language processing.

2.1 APPLICATIONS

Some applications of natural language processing are:

[1] Text Classification:
Analyze the text content and classify which category does the text belong to. Classification of email whether it belongs to phishing category of legitimate category is done by text classification.
[2] Speech Recognition:
Taking input of the sentence spoken by a person and understanding the meaning to give correct output.
[3] Machine Translation:
Translation from one human language to another.
[4] Automatic Text Summarization:
Technique to produce a short and accurate summary of long text documents.
[5] Chatbot:
Chatbot systems that are used on the websites to conduct communication with the users like answering the doubts of the users. Understanding the query and answering the same is done using NLP.
[6] Sentiment Analysis:
To identify the emotion from the posts which may or may not be expressed explicitly.
[7] Spell Checking:
Spelling of the keyword checking.
[8] Advertisement Matching:
Recommendation of ads based on the user's search history.

2.2 COMPONENTS OF NLP

In general, an application of NLP works as shown in Fig 2.

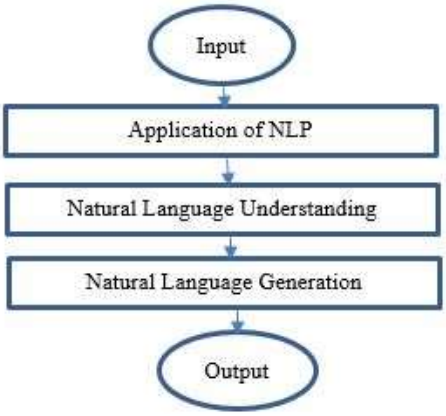

Fig 2. Components of NLP

It has two major components:

[1] Natural Language Understanding:
Taking the input from the application in the spoken or written form and deriving the meaning of the same. For example, in speech recognition application like Apple's Siri, the user speaks a sentence and Siri takes it as an input and derives the meaning by understanding it.

[2] Natural Language Generation:
Supplying the output from the internal formal representation. The planning of what to say and express in the natural language. For example, answering the question asked by the user.

III. CLASSIFICATION ALGORITHMS

Machine learning is a technique to train the system and make system take decisions of its own. There are various classifiers that are used for the training of system. They can belong to Supervised, Unsupervised or reinforcement learning. Supervised classifiers are the classifiers which are given prior training using data input and necessary output, according to which the model learns to predict and then the further classification takes place. Unsupervised learning is the type of learning where the system is given some prior knowledge about the input data, according to which model is created which has to classify the data on its own. Fig.3 shows the some of the classifiers and the category it belongs to. Reinforcement learning is the type of machine learning technique that deals with which suitable action to be taken to optimizes the reward in any specific situation. Unlike in supervised, there is no correct answer given, instead the decision is made at each task. For example training the system to play chess. The email classification study that has been carried out used the supervised or Unsupervised leaning.

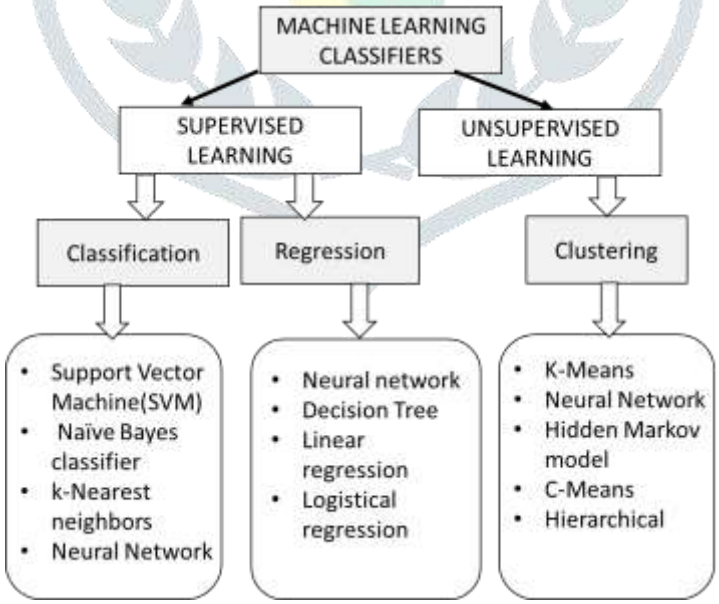

Fig 3. Machine learning classification algorithms

[1] Support Vector Machine classifier:
SVM classifier does representation of various classes in a hyperplane in n-dimensional space; n is the number of features. Then the data items are plotted and finally we draw a line/optimum hyperplane that separates the 2 classes. For example, attributes be eye color and height in a 2 dimensional space. Now, every data point will have 2 coordinates(support vectors) that are: eye color and height.

[2] Naive Bayesian classifier:
It is probabilistic classifiers. The Bayes theorem is the base of this classifier, which finds the probability of an event based on the prior knowledge of conditions. Naive Bayesian classifier assumes that in a class, the existence of one specific feature is not related to the existence of any other feature which is the naive independence.

[3] Random forest classifier:
A Random forest classifier is an ensemble algorithm. Ensemble algorithm is the combination of same or different kind of algorithms. Set of trees make a forest. Here, Random forest is a set of decision trees. The voting of each decision tree is taken and the new output case is added to that class which has highest vote.

[4] Decision tree classifier:
It is a classifier which is represented by a tree-like model or graph of decision that represents the possible results. In Decision tree we split the input data into two or more homogeneous sets. Then the decision analysis is per- formed on that data. This approach contains conditional statements. To make distinctive groups, the individual tags or attributes are applied to each node of the tree.

[5] K- nearest neighbors classifier:
KNN is a classifier which stores all the existing outputs or neighbors and then classifies a new case output by considering the neighbors. The new case output is as- signed to that class, which is most similar to its k-nearest neighbors, measured by distance function. Where k is the number of neighbors to be close to. Logistical regression classifier: Logistic model in statistics mean that event can have probability between 0 and 1 like whether a person has won or lost. Similarly, Logistical regression is a statistical classifier which is applied on a discrete binary values. For example, in spam detection, if the probability is less than 0.2 then ”less danger” else ”danger”.

IV. LITERATURE SURVEY

The Phishing email detection study has been done in many perspectives. There are various methods for the detection and prevention. Anti phishing tools, Server side, browser side methods.
Some spam identification [9] Methods :

- Blacklist and Whitelist
- Page content analysis
- Header analysis
- URL detection
- Email content analysis for keywords

Following are some of the papers about phishing attack detection approach using NLP and the outline is given in TABLE I.

S. Aggarwal et al. [1] explained an approach using Natural Language Processing. To analyze the content of the email text which are linkless. Detection of four features: Detecting absence of name, mention of money, Detecting presence of reply, sense of urgency. The score, was the combination of all 4 feature score. They used Standford CoreNLP API for NLP and RiTa Wordnet API. Dataset of 1000 emails were taken, out of which 600 Phishing and 400 legitimate emails. A static classifier was used to classify between the two classes i.e Phishing or legitimate. The score was then combined with the header analysis[10]. With this approach they[1] got the accuracy of 99.4%.

Egozi Gal et al.[2] state that many phishing email detector models previously done, ignored the features like punctuation, stopword count, Word count, uniqueness factors etc. Hence they[2] proposed a model consisting of Feature Extraction phase and Machine Learning Phase. 26 Novel Features are used here[2], including the features that were ignored previously to train the classifier models. Out of which 14 out of 17 models were tested as a success. Weighted Linear SVM was seen to the best among all. Dataset used is from IWSPA for both training and testing, for phishing emails: NAZARIO Phishing corpora and for ham emails: Wikileaks along with SpamAssassin was used. Detection of over 80% for phishing emails and 95% for ham email was identified.

Lotter et al.[3] tried to overcome the problem of human weakness of not being able to detect emails with phishing attack simply by reading the email as well as no proper software guidance. Hence they[3] proposed a framework to address the challenge which gives a user interface to help user in the detection using Rule Base filtering and Bayesian spam filtering. Detection of 10 features evaluation: urgency, personal information demand, unknown sender, fake hyperlinks, image having message as the email body, unrealistic promises, grammatical errors, signs of impersonation, malware as attachments, large set of random email addresses. The result was seen to be successful where the framework gave symbolic indication to users such as if email was safe then green, if doubtful then orange else if danger then red color.

Peng et al. [4] have presented an approach to detect phishing email attacks with the help of NLP and machine learning. To detect if content has something malicious, the intended system performs the semantic analysis of the text. An NLP technique is used to analyze each sentence and finds the semantic meaning of words with respect to the predicate. With respect to each word in the sentence, this approach recognizes whether the sentence is an order of inquiry. To generate the blacklist of malicious pairs, Supervised ML is used here. SEAHound Algorithm for detecting phishing emails and Netcraft Anti-Phishing Toolbar is used to validate whether the URL is valid or not. Implemented with Python scripts and dataset Nazario phishing email set. Results of Netcraft SEAHound are compared and obtained precision 98% and 95% respectively.

Kim et al. [5] proposes a model where the semantic analysis is carried out by focusing on the natural language text unlike the previous method having metadata i.e. header analysis and URL analysis. In this method, the attack is detected if the attacker performs one of the actions i.e. asking a private question or issuing a command like Clicking the link. A PARALEX question answer system[12] is modified, which has a database of suspicious questions with answers, is used for detection of Questions. For detection of command, verb-object blacklist is generated where verb-object pair is extracted and TF-IDF(term frequency- inverse document frequency) score is measured. This approach used 100,000 phishing emails for verb-object blacklist and 87,048 phishing emails for the testing. Precision 80% and Recall 65%.

Sahingoz et al.[6] proposed a model to detect whether the URL in the email is legitimate or phishing in real time. Features used are Word Vector, NLP based and Hybrid. To find the word vector a word list is created. Each word from the URL is separated from one another using separators, removed the digits and random letter words. Meaningful existing words are added to the list to be analyzed. Words consisting of 2 meaningful words is separated using a word decomposition module. For example SECURELOGIN is separated to SECURE and LOGIN and are added to the list. 7 classification algorithms were used namely Naive Bayes, Random Forest, KNN (n = 3), Adaboost, K-star, SMO and Decision Tree. Random Forest Algorithm proved to give the best giving 97.98% accuracy.

Babagoli et al.[7] used a nonlinear regression strategy for detecting phishing or legitimate website. The 20 most important terms were chosen among 30 features.[7] used 2 meta-heuristic-based algorithms: support vector machine(SVM) and harmony search (HS) algorithms to train the system. TO extract features decision tree (DT) and wrapper methods were used. To evaluate the features, DT was used. Out of the 2 algorithms, HS produced a better accuracy rate of 94.13% for train and 92.80% for test processes, by using 11,055 web pages as dataset.

A. Vazhayil et al. [8] proposed a new phishing email detection framework using deep learning to classify between phishing and legitimate emails. The analysis had to be done on 2 tasks: emails with header and without header. To get non sequential representation of the corpus, they[8] used Term Document Matrix (TDM). For dimensionality reduction the Singular Value Decomposition (SVD) and Nonnegative Matrix Factorization (NMF). This is incorporated to various classical machine learning algorithms like Support Vector Machine, Naive Bayes, K- Nearest Neighbors, Decision Tree, Random Forest and Logistic Regression.

IV. RESULTS

The overview of various approaches is shown in Table 1. It provides information on phishing attack detection approaches based on the number of features used for the purpose of feature extraction, the name and count of datasets, the classifier used and result in terms of accuracy or success.

Table 1. Outline of Algorithms used to detect Phishing attacks using NLP

Reference Paper	Features	Datasets	Classifiers	Result
[1]	4	600 Phishing and 400 Non phishing email set	Static	99.4%
[2]	2	IWSPA, Nazario Phishing email set	SVM out of 17 classifiers	80% Phishing and 95% legitimate accuracy.
[3]	10	Not mentioned	Rule Based, Bayesian spam filter	Successful Framework
[4]	-	Nazario Phishing email set	Natural Language Processing classifier	95% accuracy
[5]	2	187,048 phishing 100,000 and legitimate	PARALEX tool with k-best queries verb-object blacklist using TF-IDF score	Precision 80% Recall 65%.
[6]	40	NLP-based(human determined)features, 102 word vectors & 36,400 legitimate URLs 37,175 phishing URLs.	Random Forest Algorithm out of 7 classifiers	97.98% accuracy
[7]	20	11,055 Web pages	SVM and	94.13% training and

			Harmony search	92.80% testing accuracy
[8]	-	Legitimate:4082 with header,5088 non header, Phishing: 501 with header, 612 non header	Decision Tree, kNN, Logistic Regression, Naive Bayes, Random Forest and SVM	Successful framework with 4 result tables[8]

IV. CONCLUSION

Phishing attack is a dangerous cyber attack to steal personal information of the legitimate users and use the information for fraudulent act like stealing the money. There are various techniques for the detection like Natural Language Processing, Image Processing, Machine Learning, Rule-Based ,Black List or detection of an attack from a URL etc. Natural language processing for the detection of phishing attack has been proved to be successful using different approaches and models. Therefore, this study provides an in-sight to phishing attacks detection methods, study and survey on the features, classifiers that were considered and the methodology used.

IV. REFERENCES

[1] S. Aggarwal, V. Kumar, and S. D. Sudarsan, 2014, Identification and detection of phishing emails using natural language processing techniques, In Proceedings of the 7th International Conference on Security of Information and Networks, SIN '14

[2] Egozi Gal, Rakesh Verma, 2018, Phishing Email Detection Using Robust NLP Techniques, IEEE International Conference on Data Mining Workshops (ICDMW).

[3] A. Lötter and L. Futcher, 2014, A Framework to Assist Email Users in the Identification of Phishing Attacks Proceedings of the Eighth International Symposium on Human Aspects of Information Security & Assurance (HAISA)

[4] Tianrui Peng, Ian G. Harris, Yuki Sawa, 2018, Detecting Phishing Attacks Using Natural Language Processing and Machine Learning, , 12th IEEE International Conference on Semantic Computing.

[5] Kim, Myeongsoo, et al. Catch me, Yes we can!-Pwning Social Engineers using Natural Language Processing Techniques in Real-Time.

[6] Sahingoz, Ozgur Koray, et al. 2019, Machine learning based phishing detection from URLs. Expert Systems with Applications 117, 345-357.

[7] Babagoli, Mehdi, Mohammad Pourmahmood Aghababa, and Vahid Solouk. 2019, Heuristic nonlinear regression strategy for detecting phishing websites. Soft Computing 23.12 : 4315-4327.

[8] A. Vazhayil, N. B. Harikrishnan, R. Vinayakumar, and K. P. Soman, PED-ML: Phishing email detection using classical machine learning techniques, in Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal. (IWSPA), A.D.R. Verma, Ed. Tempe, AZ, USA, 2018, pp. 1-8.

[9] Malge, Amol, and S. M. Chaware. 2016, An efficient framework for spam mail detection in attachments using NLP. Int. J. Sci. Res. 5.6: 1121-1125.

[10] Fellbaum, C. (ed.): WordNet An Electronic Lexical Database. MIT Press (1998)

[11] Anti-Phishing Work Group: APWG Trends Report Q4 2016). [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf

[12] A. Fader, L. S. Zettlemoyer, and O. Etzioni. Paraphrase-driven learning for open question answering. In ACL, 2013.

[13] History of Phishing attack. Available: https://www.phishing.org/history-of-phishing