International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 333 - 340

Multi-Layer Perceptron Backpropagation for Enhanced Detection of Spam Mails

K.Padmaja[1], Nikita Manne[2], Himanshu Kumar Diwedi[3], D.Prem Kumar[4]

[1,2,3,4]*Assistant Professor*

[1,2,3,4]*Department of CSE*

[1]*S V Engineering College (SVEC), Tirupati, Andhra Pradesh*

[2]*Siddhartha Engineering College, Ibrahimpatnam, Hyderabad*

[3]*Dev Bhoomi Institute of Technology, Dehradun, Uttarakhand*

[4]*S V College of Engineering (SVCE), Tirupati, Andhra Pradesh*

[1]*padmajaskrishna@gmail.com,* [2]*sainikita@gmail.com*

[3]*himanshudiwedi01@gmail.com,* [4]*premkumar.d@svcolleges.edu.in*

Abstract

In today's technological world everything is done online, for communicating with the business partner, students, friends, etc. we used email application, which easily share files information to another user. The general problem for all the users which are used email application for communication is phishing. It is one of the types of cyber-attack in which fradaunt used the fake message and fake sites for trapping money from clients. Attackers attempt to draw online clients by persuading them to share their username, passwords, bank account information for fill billing data. One of the principle issues of phishing email location is the unknown "zero-day" phishing attack, which builds the level of trouble to recognize phishing email. These days, phishers are making diverse portrayal strategies to make unknown "zero-day" phishing email to break the barriers of those locators. There are number of techniques are developed for phishing detection, but these systems were fail to provide appropriate results. In this paper we proposed the technique for detecting the phishing mails by using multi-layer perceptron classifier. System can check the mail in the blacklist mail ids first, it email-id exist their then it can store directly, if not exist system perform the classification algorithm and detect the mail is real mail or phishing mail. System is compared with the existing system used for phishing detection.

Keywords: *email phishing, classifier, multi-layer perceptron, zero-day attack, and cyber-attack.*

1. Introduction

In a day to day life people used email application for communication, sharing files, data. It is used by government sectors, business sector, different organizations, school, colleges, and hospitals for the communication purpose [10]. There is one issue called as phishing email which is faced by the user while handling the email application. The term phishing means a subset of spam which is identified with social engineering plans, which relies upon forged mails and after that through an embedded links in the email, the phisher tries to divert clients to fake sites. These websites are intended to get money related information from the user who visits their links fraud, including usernames, passwords, and credit card numbers. Occasionally, the phisher tries to mislead the client to a fake site or to a legitimate one checked by intermediaries [11].

The issue of phishing email is winding up more terrible. One of the reviews on phishing attack reported that on phishing assault demonstrated that around 3.6 million clients in the USA lose money in light of phishing [12]. As per the survey in 2007 the total loss amount is evaluated at US$3.2 billion dollars also the number of persons expanded from 2.3 million to 3.6 million. Latest report by eCrime trends report clarify that phishing attacks has expanded by 12% in every year. Numerous issues emerge because of phishing emails, the majority of which influence financial organizations also, their

customers. Phishing is fit for harming electronic trade since it makes clients lose their trust on the Internet. One of the fundamental issues in email is unknown 'zero-day' phishing emails. Zero-day attacks are characterized as attacks that phishers mount utilizing has that are most certainly not boycotted or utilizing methods that avoid known methodologies in phishing recognition [13]. Phishing email is complex to the point that it can't be identified by a large number of current methods in light of the fact that the phisher can utilize new vulnerabilities which are never observed before. There is various conceivable solution for these issues, however not successful to solve these issues. These reaches from correspondence arranged methodologies like verification conventions over blacklisting to content based filtering approaches which for the most part rely upon some of Artificial Intelligence strategies. Current AI techniques are capable to distinguish phishing email in light of fixed features and standards while a couple of number of machine learning protocols configuration to work in online mode. The level of errors in the classification process will increment after some time, particularly when managing with obscure zero-day phishing emails.

There are number of studies done on phishing email detection, some them are failed to detect zero-day phishing email. In this paper we proposed the technique which detects the phishing mail by using the multilayer perceptron classifier. These systems have the capability of determining whether the mail is phishing or harm. The main goal of this paper is to propose the system for detecting the phishing mails by using the classification technique. Also proposed the system for solve the issue of detecting the zero-day phishing mails [19][20]. The objectives of the systems are:

1. To develop a method with low computational cost and to analyses its performance under different scenarios.
2. To improve the performance and accuracy in terms of the classification and prediction of phishing e-mail in the future.
3. To optimize memory consuming in the classifier process and to reduce the time needed for classify the email with unlimited learning, while the characteristics of phishing e-mail features have change.
4. To evaluate the proposed framework against using approaches for the purpose of phishing email detection.

2. Literature Survey

In [1] an efficient review examination on existing works related with the phishing identification and response methods together with apoptosis have been additionally researched and evaluated. Besides, one contextual analysis to demonstrate the confirmation of idea how the phishing functions is additionally examined in this paper. This paper additionally discusses about the difficulties and the potential research for future work related with the integration of phishing detection model and reaction with apoptosis. This paper additionally can be utilized as a kind of perspective and direction for further study on phishing detection and reaction.

A computer executed technique for profiling cyber threats identified in an objective domain, including: receiving from a Security Information and Event Manager (SIEM) checking the objective condition, alerts activated by a detected potential cyber threat and, for each alert; recovering captured packet information identified with the alert; separating information relating to the set of attributes from captured protocol information triggered the alert; applying fuzzy logic to information relating to at least one of the attribute to decide esteems for at least one yield factors demonstrative of a level of a part of risk attractable from the cyber threat[2].

In [3] actualized a desktop application called PhishShield, which focuses on URL and website content of phishing page. Phishshield takes URL as input and output the status of URL as phishing or honest to goodness website. The heuristics used to distinguish phishing are footer joins with null values, zero connections in group of html, copyright content, title substance and site character. PhishShield can identify zero-hour phishing attacks which boycotts unfit to recognize and it is speedier than visual based evaluation methods that are utilized as a part of detecting phishing.

In [4] proposed a novel hybrid phish detection method based on phishing boycotts and phishing properties. Author used some new phish from PhishTank that were recently added to test that it can be detected by blacklist or not. Author found that 70 % of the phishing websites in their dataset lasted less than two hours. Boycotts were ineffective when protecting users initially, as most of them caught less than 20% of phish at zero hour.

In [5] presents a novel hybrid framework that coordinates irregularity, conduct and signature-based systems for recognizing and analyzing zero-day attacks progressively. It has layered and modular outline which accomplishes superior, adaptability and versatility. The framework is implemented and

assessed against different standard measurements like true positive rate (TPR), false positive rate (FPR), F-measure, total accuracy (ACC) and receiver operating characteristic (ROC) curve. The outcome indicates high location rate with almost zero false positives.

In [6] propose a phishing discovery and counteractive approach combining URL-based and Webpage similitude-based identification. URL-based phishing recognition includes extraction of actual URL and the visual URL. Link Guard algorithm is utilized to examine the two URLs lastly relying upon the outcome created by the protocol the method continues to the next stage. In the event that phishing isn't distinguished or phishing possibility is anticipated in URL-based identification, the algorithm continues to the visual comparability-based recognition. A novel method to outwardly compare a suspicious page and the legitimate one is displayed.

In [7] propose a security threat grouping model which enables author to consider the threats class affect rather than a risk affect as a risk varies after some time. This paper tends to various criteria of data framework security risks classification and gives a survey of most threats' classification models. They characterize a hybrid demonstrate for data framework security threat characterization with a specific end goal to propose a classification design that supports all risk order standards and enables associations to execute their data security techniques.

In [8] proposed a hybrid solution for protect against zero-day phishing attacks. In the proposed approach, the idea of coordinating each URL with trusted domains is utilized from the Link Guard protocol and the idea of CSS coordinating is utilized from the BaitAlarm scheme. Proposed approach is productive and covers an extensive variety of sites phishing attacks, and result less false positive rate.

In [9] manage the above issue by proposing an AC calculation called enhanced multi-label classifiers based associative classification (eMCAC). This protocol finds rules related with a set of classes from single label data that other current AC protocol can't prompt. Moreover, eMCAC limits the quantity of extracted rules utilizing a classifier building technique. The proposed calculation has been tested on a genuine application dataset identified with site phishing and the outcomes uncover that eMCAC's precision is exceptionally competitive if contrasted with other known AC and great characterization protocol in information mining.

3. Proposed System

3.1 System Overview

In this system we are using different emails received by the user as an input dataset. From this emails training file is generated and stored in the database. After that the by entering the new mail system initially check that the whether this mail is from blacklisted mail ids if not then some features are extracted from the mail and finally classification process is done for detecting whether it is phishing mail or not. For classification multi-layer perceptron classifier is used. The architectural view of the system is as follows

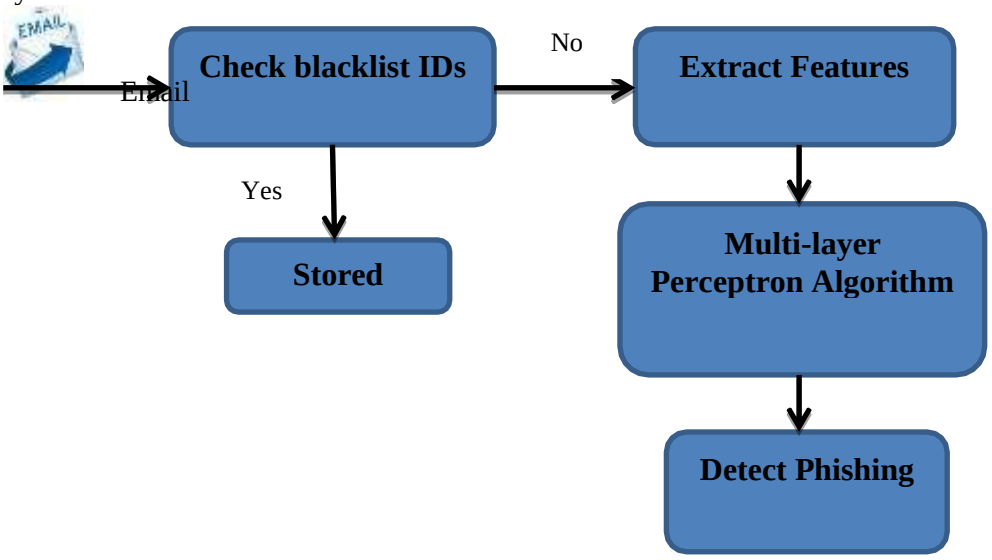

Fig 1. System Architecture

3.2 Multi-Layer Perceptron Algorithm

Input: Query
Dataset: all friends information / existing dataset + real time data collected from friend list
Dataset: Matched FriendList[]=null;
Process:
1 Create test file using dataset.
2 Take query as input & creating training file related to query.
3 Pass training & testing file to multilayer perceptron.
4 Matching queries on test data assigning
class 5 If query match then
6 class =1 // Phishing email /spam email
7 else
8 class =0 // legitimate email /ham Email
9 Stored emails who has class = 1 in matched email id list
10 End

4. Result and Discussion

4.1 Technology Used

We used java technology for implementing the system with netbeans IDE. Different library files are used for implementation. We are not used any specific hardware for this system.

Dataset Description

From email datasets email subject, email body, email signature, and link embedded in emails, sender, receiver name, sender and receiver are known and unknown, use of word in email update, urgent, click these features are mainly considered.

This dataset is composed of a selection of mail messages, suitable for use in testing spam filtering systems.

4.2 THE DATASET CONTAINS TWO PARTS:

- TRAINING: 4327 messages out of which there are 2949 non-spam messages (HAM) and 1378 spam messages (SPAM), all received from non-spam-trap sources. SPAM Train label contains the labels of the emails, with 1 stands for a HAM and 0 stands for a SPAM.
- TESTING: 4292 messages

4. Result and Discussion

4.1 Technology Used

We used java technology for implementing the system with netbeans IDE. Different library files are used for implementation. We are not used any specific hardware for this system.

Dataset Description

From email datasets email subject, email body, email signature, and link embedded in emails, sender, receiver name, sender and receiver are known and unknown, use of word in email update, urgent, click these features are mainly considered.

This dataset is composed of a selection of mail messages, suitable for use in testing spam filtering systems.

4.2 THE DATASET CONTAINS TWO PARTS:

- TRAINING: 4327 messages out of which there are 2949 non-spam messages (HAM) and 1378 spam messages (SPAM), all received from non-spam-trap sources. SPAM Train label contains the labels of the emails, with 1 stands for a HAM and 0 stands for a SPAM.
- TESTING: 4292 messages

4.3 Experimental Setup

In this experimental setup, we divide into three parts, experiment 1 is called as before training approach, actual training and lastly after training approach. In this experiment mainly we want to provide user education though online training about phishing emails.

4.3.1 Experiment 1 (Before Training)

According to [14,15] the following groups are more susceptible to phishing: (1) age group 18 to 25 years, (2) those who do not have a computer education and (3) females with less technical knowledge. For our survey therefore we chose final year and graduate students of computer engineering/IT. The number of participants was 149. In experiment 1 (before training) only 3 legitimate and 13 phishing emails were used. Because our motto is to identify phishing emails by users. This experiment we selected According register media phishing categories as 5 types which are shown in table 1, with

percentage of year 2016.also for training total examples for each category also given. In experiment 1 we shown 16 emails, which is really received on authors emails... we selected participants according to groups 1, and 2 mention and in 5.1 that tis computer engineering final year students. This participant are from age group in range from 18 years to 25 years and they are having computer knowledge then we gather all participant s those are willingly interested to participant in this experiment and explain all 3 experiments in depth with motivation. then we show to participants all 16 emails to identification. Also, we are provided answer sheet hard copy with question numbers
. while showing email we told participant to write answer as three options which is fixed like option A is phishing. option B legitimate, and option C.
We show 16 emails out of that 3 are legitimate and 13 are phishing emails. After showing emails we told participant to write down answer as mentions above.

Table1. Phishing Email Categories (No. Of Email Example For Survey 2&3) [18]

Phishing email categories	Total no of emails (141)	Percentage (%)	Email example assume 9 is minimum value in categories=1 example
Health care insurance investment (49+13)	62	43.97%	7
Finance, credit card, loan, scholarship, win prize, offers (18+15+12)	45	31.91%	5
Authentication of email and bank account	14	9.92%	2
Job offers	11	7.80%	1
IT	9	6.38%	1
Total	141	99.94%	16

4.3.2 Experiment 2 (Actual Training for Identification of Phishing Emails)

In this experiment we continue all participant as it is for training first, we discuss all emails, which is shown in before training ask answer orally and explain why particular email is phishing or legitimate, also how you are going to identify email is legitimate and phishing. Also does and don't. we give phishing email identification tips like to check sender is unknown, observed email header, email subject, email footer, email language, multi-color in email, email is embedded with link or not if yes then weather asking personal information like bank details and user name and password. Also, some online email [16, 17] examples are discussed.

4.3.3 Experiment 3(After Phishing Email Identification Training)

In experiment participant are very excited to identification of emails. To check confidence level of participant on scale of 5. We summarized and consolidated and generated graph for before and after training confidence level of participant, which is shown in fig.2. For experiment 3 we selected 16 emails according to table1 also we repeated 40% email of before training as it is and 60% emails are newly introduced. Experiment 1 and experiment 3 participants responses are notes and compared which is shown in result as next section.

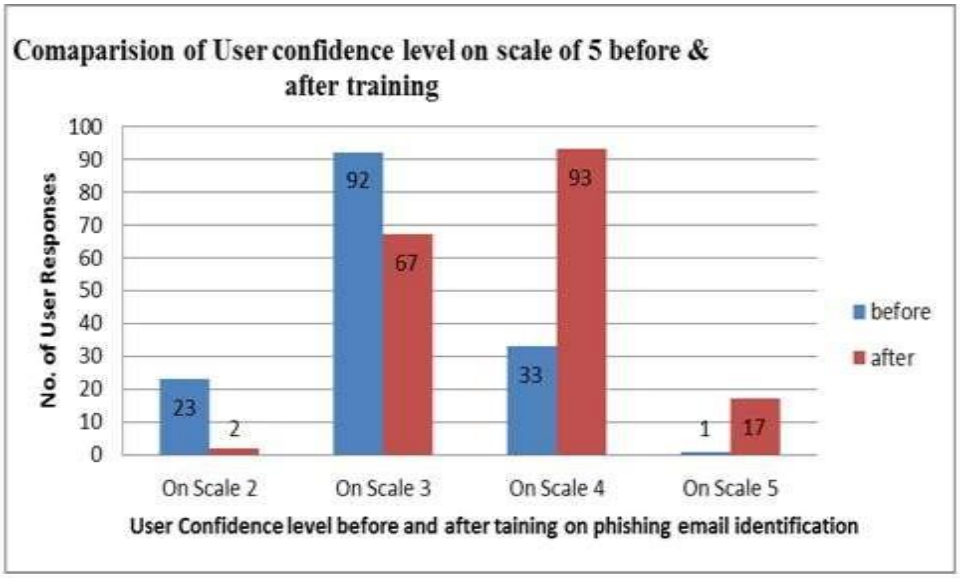

Fig.1 Confidence level of participant before and after training [18]

We ask participant confidence level 1 and 3 on scale of 5. We summarized and consolidated and generated graph for before and after training confidence level of participant, which is shown in fig.2. In experiment 1 92 participant confidence levels is 3.but after training 67 participant confidence levels is 3 and 93 participant confidence level are 4. So, it indicates that after training confidence level of identification of phishing meal is increases only 49%before training and 38% after training legitimate emails are classify by participant which is less. Before training 42% and after training 59% phishing emails are identified.it indicated that after training phishing email identification is increased by 17%. As before training approach phishing and legitimate email correctly identification performance is less than 50%. So that we apply training approach for same participant we repetition of 40% email example of before training and see performance of users. In training all example which is included in before training with do's and don'ts are explain also phishing identification tips are given.

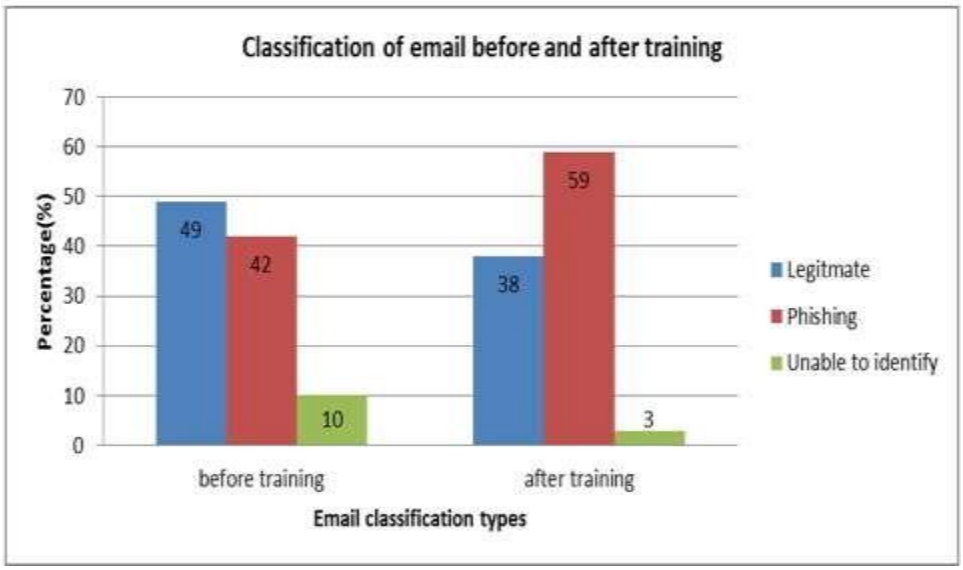

Fig.3 Before training legitimate email correctly and incorrectly classify in percentage.[18]

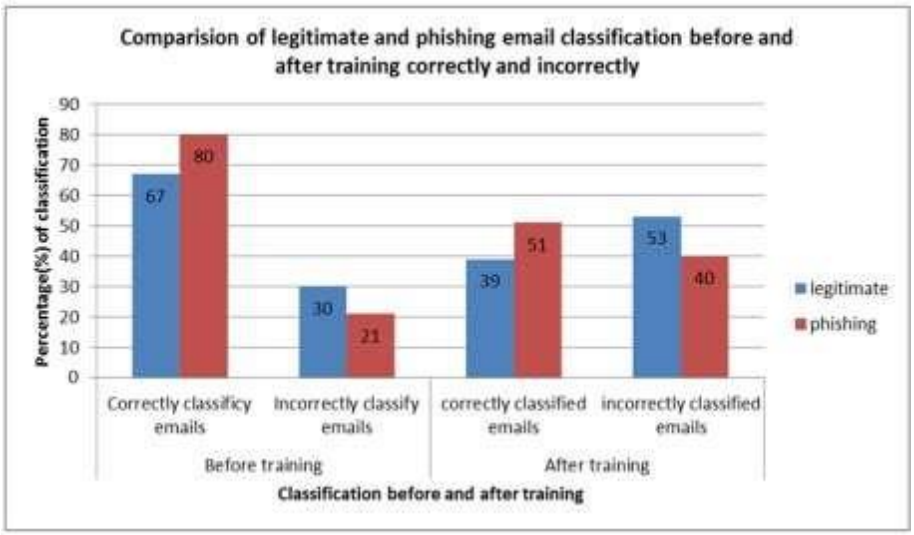

Fig.4 comparison of Phishing email classified as phishing correctly before and after training [18]

After training 67 % users correctly identify legitimate email and 80 % phishing emails are identified. If we compare before and after training approach only 28% users' legitimate email correctly identification is improvement and 39% phishing email identification improvement, which is very less so that we required to solve this problem machine learning algorithms are required. After training we take review of users why they incorrectly classify legitimate email as phishing and phishing email as legitimate. They give reason like multicolor are used in email, email embedded URL is given, sender is unknown, email signature is not proper, domain and subdomain is not register**.**

5 Conclusion

In this paper discussed the work proposed for detecting phishing in emails. We used multi-layer perceptron algorithm for detecting the mail is phishing or real mail. These frameworks detect and predict the zero-day phishing email with decreasing the level of false positive rate of phishing emails. This system improves the accuracy also increase the performance of the classification and prediction of phishing email values. System finally compared with the existing systems and proves that the system performs better than the existing framework.

References

1. Yahaya, Saudi, Madihah ,Abdullah, Ismail. "A Review and Proof of Concept for Phishing Scam Detection and Response using Apoptosis". International Journal of Advanced Computer Science and Applications (IJACSA) ,Volume 8,Issue 6.
2. S .Laidlaw and M .Hillick, "Profiling cyber threats detected in a target environment and automatically generating one or more rule bases for an expert system usable to profile cyber threats detected in a target environment". U.S. Patent 9,503,472. Cyberlytic Limited, 2016.
3. Routhu Srinivasa Rao∗ and Syed Taqi Ali, "PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach", Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015). Volume 54, Pages 147-156
4. Namrata Singh, Nihar Ranjan Roy, "A Hybrid Approach to Detect Zero Day Phishing Websites", International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1761-1770
5. Ratinder Kaur and Maninder Singh, "A Hybrid Real-time Zero-day Attack Detection and Analysis System", I. J. Computer Network and Information Security, 2015, Volume 9, 19-31.
6. N. M. Shekokar, C. Shah, M. Mahajan, S. Rachh, "An Ideal Approach for Detection and Prevention of Phishing Attacks", Procedia Computer Science Volumn 49 (2015) page no. 82 – 91.

7. Mouna Jouini, Latifa Ben Arfa Rabai, Anis Ben Aissa, “Classification of security threats in information systems”, 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014).Volumn 32,page no. 489-496
8. A. Mishr and B. B. Gupta, “Hybrid Solution to Detect and Filter Zero-day Phishing Attacks”, Emerging research in computing, information, communication and applications, ERCICA 2014. page no.373-379
9. Neda Abdelhamid, “Multi-label rules for phishing classification”, Applied Computing and Informatics (2015) volume 11, page no.29–46.
10. MAAWG (2011). Messaging Anti-Abuse Working Group (MAAWG) Email Metrics Program. 15. third Quarter.
11. Singh, D. K., & Ashraf, M. (2019). Detect the phishing websites in the contex of internet security by using machine learning approach. International Journal of Advanced Science and Technology, 27(1), 104-111.
12. APWG (2010). “Phishing Activity Trends Report“.From http://www.antiphishing.org/reports/apwg_report_Q1_2010.pdf.
13. GARTNER (2007). “Gartner Survey Shows Phishing Attacks Escalated in 2007; More than $3 Billion Lost to These Attacks.” Retrieved December 17,from http://www.gartner.com/it/page.jsp?id=565125.
14. Bimal Parmar, F. (2012). “Protecting against spear-phishing.” Computer Fraud & Security 2012(1): page no 8-11.
15. Steve Sheng,1 Mandy, Holbrook, Ponnurangam Kumaraguru, Lorrie Cranor,Julie Downs, “Who Falls for Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions”, Copyright 2010 ACM.
16. K. Manoj, T. S. Sandeep, N. Sudhakar Reddy and P. M. D. Alikhan, "Genuine ratings for mobile apps with the support of authenticated users’ reviews," 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), Bangalore, India, 2018, pp. 217-221.
17. T. S. Sandeep, K. Manoj, N. S. Reddy and R. R. Kumar, "Big Data Ensure Homologous Patient Enduring Therapy Time Forecast Algorithm by Healing Facility Echelon Recommendation," 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), Bangalore, India, 2018, pp. 320-325.https://blog.returnpath.com/10-tips-on-how-to-identifya-phishing-or-spoofing- email
18. http://www.techrepublic.com/blog/10-things/10-tipsfor-spotting-a-phishing-email
19. Dhamdhere V., P. Joeg ,” To Study of Phishing Attacks and User Behavior”, International Conference on Inventive Computation Technologies (ICICT 2017)
20. Dhamdhere V., P. Joeg ,” A Study User Behavior Using Phishing Education and Training”, International Journal of Engineering Research in Computer Science and Engineering,2017,pp. 50-55.
21. Dhamdhere V, S. Vanjale,” A novel approach for phishing email real time classification using k-mean algorithm, International Journal of Electrical and Computer Engineering,2018,pp.5326-5332.
22. Dhamdhere V, S. Vanjale,”PHISH SAFE GUARD-Phishing Detection: Enhance Anti-Phishing System Using Machine Learning Algorithm”, International Journal of Engineering and Advanced Technology, 2019, pp. 1668-1671.
23. K.Manoj Kumar, T.S.Sandeep, G.Sunil Kumar, K.Anusha, "Enhanced Text Mining Methodology in Social Media Platform," International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 8, no. 12, pp. 4857-4861, 2019.