International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Detecting Forged E-Mail using Data Mining Techniques

Prasanta Kumar Sahoo, Cheguri Rajitha

Abstract: ***In the modern era of computers E-mails is becoming a very important mode of communication for industry, people, and organizations and for the society as a whole. Especially in corporate sectors and business organizations, E-mails are widely used for business and personal communication. The feature of E-mails is that, it creates quick, reliable type of communication that's all free and simply accessible. In spite of so many alternative means of communication such as messages, social networks like WhatsApp, Twitter, and mobile applications, the uses of E-mails continuously growing exponentially. But due to its popularity there are continuously threats and attacks are carried out over E-mails for various gain. The foremost popular attack over the web is phishing mails. Phishers utilize E-mail services quite expeditiously in spite of different detection and hindrance techniques already in situ. Most of the present day phishing attacks use E-mail as the primary carrier. Phishers conceive to fraudulently acquire sensitive information, like usernames, passwords and master card details, by masquerading as a trustworthy entity in transmission. Even though there are a lot of existing techniques offered to notice phishing attacks, every one of them have their own limitations. This research aims to identify the phishing E-mails using classification techniques with a better accuracy. The technique proposed in this research work to classify forged E-mails from the Genuine E-mails and also examines the effectiveness of detection of common user's phishing E-mails. It provides a great help to the common man by proper detection of phishing attacks and protecting their confidential data.***

Key Words: Phishing attack, Fake E-mails, data mining, anti-phishing techniques.

I. INTRODUCTION

E-mail is changing into the foremost convenient means for exchanging messages electronically between one individual and another across the globe. It has become a really necessary communication medium for industry, people, organizations, and also for the society. Particularly in collective sectors and business organizations, E-mails are extensively used for business communication. On the technical aspect, it involves variety of protocols, like SMTP, POP, TCP /IP then on, for taking messages from one mailbox to a different. E-mail has been capably described as a technique for exchanging digital data from one sender to at least one or number of recipients and it has become the quality medium of communication in numerous areas of life.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Dr. Prasanta Kumar Sahoo*, Professor, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana, India. E-mail- prasantakumars@sreenidhi.edu.in

Cheguri Rajitha, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana, India. E-mail- rajithacheguri1407@gmail.com

It provides several enticing options by its virtue like quick, easy and free access, world acceptance. E-mail becomes essential because of being the key for an oversized variety of web services like social networking, news report subscriptions, etc. Here, within the initial section, we tend to introduce regarding the conception of E-mails. Then we approach techniques to encounter E-mail attacks within the second section. Within the third section, we tend to place the approach in our planned methodology that the user will apply to safeguard themselves from E-mail attacks. The various uses of E-mail are defined below:

A. ***A standard means for inter communication***: As a medium of communication, one of the most necessary functions served by E-mail is that of eliminating distances and serving to stay connected. Be it family or recent friends or simply concerning anyone, with E-mail no one is ever away. Gone the times once obtaining replies letters meant weeks of waiting amount. With E-mail, one will get a reply instantly and may even communicate in getting ready to time period for college students staying far from their oldsters, nothing will function higher as an E-mail will during this manner, it becomes a general methodology of communication for everybody.

B. ***Educational Purpose:*** In many Institutions, E-mail acts as a medium to share the course material. Lectures send course material and relevant course files or documents through E-mail to all the students in major universities across the globe. This makes it probable to receive study material and complete the course on time. The facility of attaching files opens an option also whereby study material such as course files documents and relevant videos can be sent seamlessly over E-mails within a matter of seconds.

C. ***Corporate Usage*****:** E-mail is typically advised as the official medium of communication in the business sector. Nowadays, liberal foreign policy is turning into a lot of and a lot of acceptable and therefore, the export and import of products to and from the foreign markets have become a frequent affair. This is often wherever E-mail comes into the image containing business and trade. Nowadays, most of the business-related communications are done via E-mails. All totally different forms of deals, tenders, quotations, are communicated through E-mails.

D. ***E-mail and its utilities:*** E-mails could be a store and forward service wherever it's not necessary for the receiver to be gift at the pc. The message resides within the receiver's mailbox till it is browsed or deleted.

File data, sound, image, audio or video clip also can be sent through E-mail as an attachment. Services offered by Facebook, Blogger, marital sites like bharatmatrimony.com, shaadi.com, instructional services like knowafest.com, pratimahasabha.com, etc.

Phishing has become the one of the major type of web fraud which depends on tricking clients to reveal their own information just as passwords and MasterCard details. This phishing attack is normally done by E-mail. A case of E-mail Phishing: As if E-mail seems to be from eminent sites, from a client's bank, MasterCard Company, E-mail, or net administration provider. Generally personal data like MasterCard detail, User ID are approached to update. These messages contain uniform resource locator (URL) interface that guides clients to an alternate site. This site is extremely an imagine or changed site. When clients go to this site, they're approached to enter individual information to be sent to the phishing assailant [1, 2]. Phishing is regularly used to become familiar with somebody's secret phrase or charge card data. With the assistance of E-mail arranged as though originating from a bank or authority establishment, PC clients are coordinated to phony destinations. All in all, the data that is stolen by a phishing assault is as per the following:

- User's Identification number
- User PIN and client username
- Master card details
- Online banking data

Phishers utilize various systems to trick the clients and take their own and monetary data. Scorned E-message has been created as forged sites which are most utilised by phishers. Messages which are sent by the attackers are sent to the customers asking their own information by altering few bits of the message and making the customer trust it to be from an authenticated origin. Internet spoofing is a procedure where a Fake site that seems to be indistinguishable from the original one is made by the assailants to acquire clients to take their data.

Forged E-mails frequently look incredibly authentic, and the genuine sites where the Internet client is approached to enter individual data which seems to be appears similar with the genuine one [3]. Phishing messages proliferate over E-mail, SMS, moment flag-bearers, long range interpersonal communication locales, VoIP, etc, yet E-mail is the famous method to play out this assault and 65 per cent of the all phishing assault is accomplished by visiting the hyperlink which are attached with the E-mail [4].

Numerous scientists have examined phishing assault issues, and numerous arrangements has been been suggested to distinguish spoofing assaults at various stages. Few models identify spoofing assault on the website [5, 6], further is at the E-mail level[7]. At the website level, the assurance stage checks whether the client attempting to open the site is authentic or forged. The subsequent stage will work on distinguishing spoofing assaults on the Initial level or E-mail level. Attackers any how convince to the clients to redirect to click on forged sites and this well-known procedure used by them. Therefore the model worked on distinguishes spoofing assaults in the E-mail stage, which is the main procedure that discovers at site page level, for the accompanying reasons. Right off the bat, it won't hinder the route of sites, where it ordinarily begin the identification procedure when the client click on hyperlink and the program isn't showing the ideal site until its distinguished as a genuine site. Besides, it gives a progressively assured condition to clients. Example, As soon as the client clicks on the website it immediately downloads few secret on the client system. At long last, the normal duration of a forged web site is just around 2.25 days [8].

II. REVIEW OF THE CASE STUDIES.

A. Case Study: Website Phishing Experiment

Here could be contextual analysis inside which, The site was a sure copy of the underlying Jordan Ahli Bank site www.ahlionline.com.jo intended to draw clients and instigate them by focused phishing messages to present their qualifications (username and secret phrase). The example was far reaching of our associates at Jordan Ahli Bank when accomplishing the required approvals from our administration. We will in general intentionally place different world renowned phishing choices and factors once making the faked site to leave the client's consciousness of those sorts of hazard. For example, utilizing IP address as opposed to space name, http rather than https, poor plan, composing framework blunders, nonappearance of SSL lock symbol and fake security declaration. We will in general focus on a hundred and twenty staff with our misleading phishing E-mail, educating them that their e-banking records are at the shot of being hacked and mentioned them to sign into their record through phony connection associated with our E-mail utilizing their standard customer ID and secret phrase to check their equalization at that point close more often than not. As appeared table two, the site effectively pulled in fifty-two out of the 120-focused on staff speaking to four hundred and forty yards, UN office collaborated totally by following the misleading headings and presenting their real accreditations (client ID, Password). incredibly IT division staff and IT evaluates on going eight out of the hundred and twenty unfortunate casualties speaking to seven-membered, since we will in general anticipate that them should be extra alarm than others. From elective offices forty-four staff individuals from the 120-focused on representative's unfortunate casualties speaking to thirty-seventh, fell into the bait and presented their accreditations with none dithering. The staying sixty-eight out of a hundred and twenty speaking to fifty-six was isolated as pursues: twenty-eight staff individuals prepared mistaken information that hopes to point a careful interest speaking to 23%; and forty staff, got the E-mail anyway neglected to react in any regard speaking to thirty-third. The outcomes unmistakably demonstrate that emphasis on phishing issue is uncommonly hazardous since essentially a large portion of the staff UN office reacted were defrauded; fundamentally, prepared staff like those of IT Department, and IT Auditors. Expanding the consciousness of all clients of e-banking identifying with this hazard issue is incredibly proposed. [9]

B. Case Study: Phone Phishing Experiment

For our testing example, a gathering of fifty staff was reached by ladylike associates apportioned to draw them into utilizing their own e-banking accounts, client names and passwords (through social ,and inviting discussions in view of a tricky reason). The outcomes were incredibly past desires; Most of the staff fell for the trap. When directing amicable discussions with them for a couple of times, our group figured out how to tempt them into utilizing their web banking qualifications for phony reasons. A considerable lot of these weak reasons encased checking their benefits and openness, or checking the record's honesty and property with the net server for support capacities, account security and protection affirmation. To guarantee the validity of our solicitation and to allow it a social dimensional pattern, our group needed to get in touch with them over and over, possibly 3 or multiple times. The group figured out how to mislead sixteen out of the fifty staff into utilizing their full e-banking certifications (client name and secret key), that delineated thirty-second of the example. This extent is considered a high especially once we comprehend that the exploited people were the specialists individuals from a bank, World Health Organization square measure assumed to be amazingly taught with pertinence the dangers identified with electronic financial administrations. A total of eight staff individuals (16%) joined to concede their client name just and abstained from utilizing their passwords under any conditions regardless of the reason. The staying fifty-two (26 representatives) were horrendously mindful and declined to uncover any information identifying with their accreditations via telephone. A synopsis of the outcomes uncovers the high danger of the social building security issue. Social designing comprises prompt interior danger to e-banking internet providers since it hacks legitimately and inside into the records of e-bank clients. The outcomes moreover demonstrate the desperate must be constrained to build the notice of customers not to fall casualties of this kind of risk which may have crushing results.[10]

C. Case Study: Business E-mail Compromise (BEC)

In 2017, the Nigeria-based Business E-mail Compromise (BEC) attack hit over fifty countries, targeting over five hundred businesses, preponderantly industrial corporations. The phishing scam prompted recipients to download a malicious file. Once the file was downloaded, malware would gain licensed access to business information and networks [11].

D. Case Study: Shipping Information

In July 2018, internet security cpany Comodo disclosed a new type of phishing scam specifically targeting small businesses. Phishing E-mails were sent out to more than 3,000 businesses, including the subject line 'Shipping Information'.

The E-mail noted a forthcoming delivery by United Parcel Service (UPS) and included a seemingly innocent package tracking link. When the recipient clicked on the link it contained malware, potentially releasing a virus [11].

E. Case Study: USA, Forcellina Case (2004)

A Person of age23, accessed chat rooms, used device to capture screen names of chat room participants; then sent E-mails pretending to be ISP requiring correct billing information, including current credit-card number. Used credit-card numbers and other personal data to arrange for wire transfers of funds via Western Union, but had others pick up funds from Western Union [12].

III. EXISTING WORK

Tan et al. [13] proposed a method for anti-phishing technique in which the major element of the URL is selected, for example, Meta, title, body labels. He has focused most of the data which is on left side of the URL as opposed to the right side of the URL is completed in light of the fact that the attacker endeavour to emulate the phished site as the genuine site. The whole URL is splited into token of words in which the original keywords are compared with search engine after that matching is performed The first space name and the given area name are coordinated additionally with the nation code area. On the off chance that the nearness of page equivalent with nation code which is on top area it is taken as certified website page else it is forged site page.

Yan et al. [14] proposed an attempt on Chinese phishing on Online shopping sites. The highlights utilized for the location of phishing are URL and the internet highlights and successive negligible advancement calculation. To enhance the highlights parameters they have utilized hereditary calculation. Web ZIP instrument is utilized for gathering and highlight extraction from the source code of the E-trade site page. The information mining instrument weka is utilized on preparing future framework.

Li et al. [15] proposed an AI approach for the discovery of phishing site pages. This paper accentuations on highlights of the site page, for example, web picture and report item model to advance the highlights that are separated from the site page they have utilized quantum motivated transformative calculation. The streamlined highlights are sent into weka tool help support vector machine to group the website as genuine or forged.

Liu P et al. [16] had attempted to locate a viable answer for sifting spam messages in their work. The prescribed methodology for the investigating of the utilized content is about the E-mail as a watchword just to execute multiplex word handling. In their investigation led, 4327messages in the CSDMC2010 SPAM preparing informational collection were assessed. The outcome of the demonstrated models shows an exactness of 92.8 per cent.

Thomas J. et. al. [17] has depicted property determination systems utilized in the conventional content grouping for spam separating in this study. For example, E-mail and the subject body. Diverse component determination techniques are displayed similarly. Because of broad investigations, E-mail header and body have been demonstrated to be viable for E-mail characterization.

An alternate way to deal with the recognition of phishing messages was proposed [18]. Bayes calculation utilized as an arrangement calculation with half and half highlights that consolidate content-based and conduct based highlights. The half breed highlights are, right off the bat, which are in E-mail header, for example, subject data , sent data, as well as conduct based data. Also, the body-based data includes: (a) Link-based. (b) Access key-Based. (c) Numerical number-based. (d) Manuscript-Based data. The exactness of 96 per cent was accomplished, with the 4 per cent false negative and false positive rate separately. The outcome of this works is established by the high false positive rate and exactness.

IV. PROPOSED SYSTEM

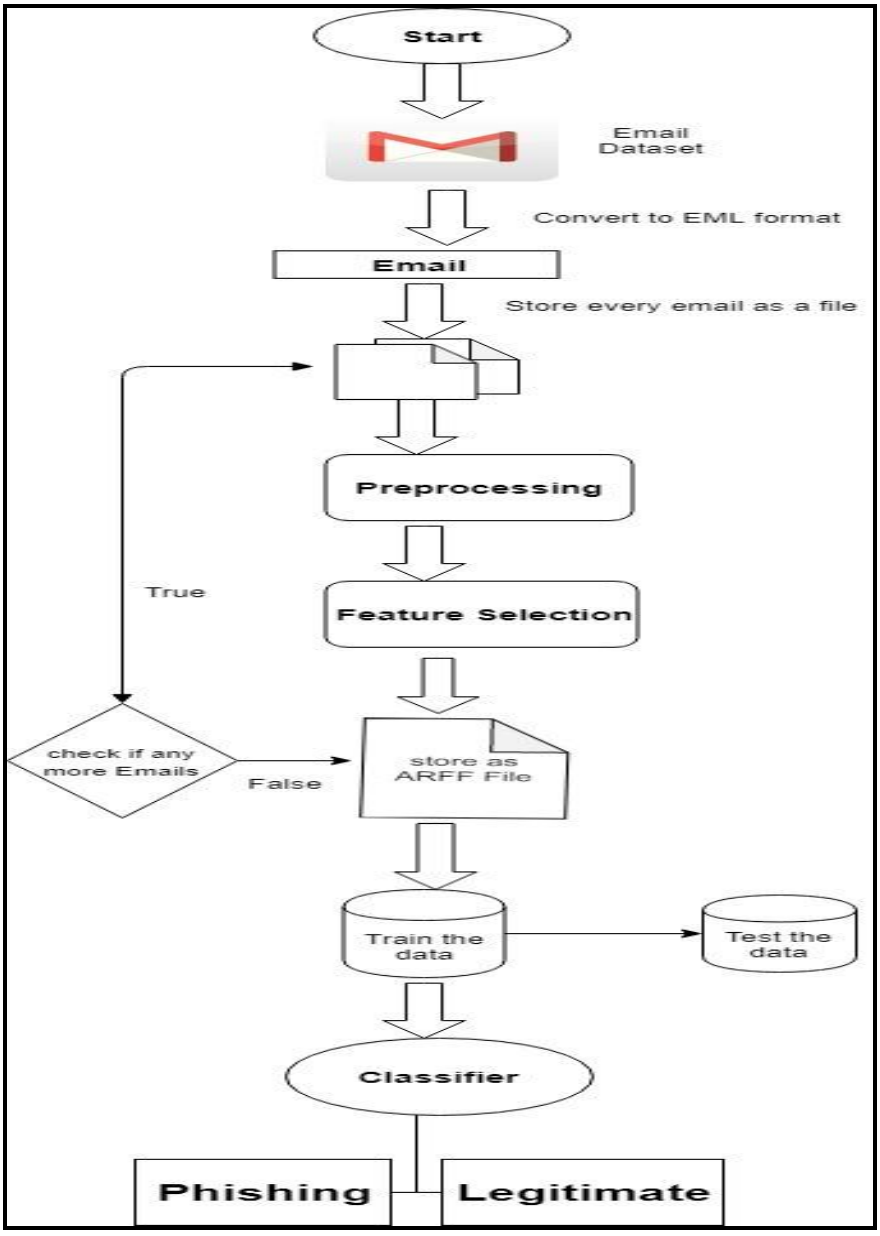

Fig 1: shows the architectural model of the proposed system.

As shown above in figure 1, the architectural model of the proposed work is to differentiating the fake E-mails from the Genuine E-mails using classification techniques. The details are given below step wise:

1. The primary step in building the E-mail classification model is by selecting the appropriate E-mail Dataset. The E-mails dataset includes both the original mails and also the fake mails.

2. Once data is collected, spilt each and every E-mail and convert them into EML (Electronic Mail Format) format, EML files are used to store every message as one file and attachments.

3. Apply data Pre-processing technique to the EML (Electronic Mail Format), as a result E-mail is splited into a token of words which is called as tokenization. It also eliminates unnecessary words and stop words to reduce the number of data that need to be examined in the pre-processing step to maximize the efficiency.

4. Once the pre-processing step is done, the result of that is going to be the input for feature selection process. Then the features such as body of the E-mail, header of the E-mail, URL, To, From, Cc, Bcc, etc are extracted from the dataset.

This process continuous until and unless all the E-mails are scanned properly from the data set and features are extracted from them.

5. The outcome of the fourth step needs to be converted into ARFF (Attribute Relation File Format) so that the classification algorithm can be applied to it. This paper proposed to use J48 classifier for E-mail dataset classification. Decision tree J48 which is the extension of ID3 (Iterative Dichotomise 3), that creates a classification model which predicts the value of an attributes (often referred as classes and instance) based on the given input attributes. It can also handle E-mail datasets with errors, missing values and continuous attribute values.

6. Based on the given attributes, the classifier divides the Datasets into Training data and Testing Data.

7. A classifier is made based on the rule, and also the features are chosen.

8. The ordered model indicates the E-mail as a phishing E-mail or the real E-mail contingent upon the precision of the model.

V. RESULTS AND DISCUSSION

The proposed architectural model is tested using the Enron dataset; The E-mails dataset includes both the original mails and also the fake mails. At first Data Pre-processing is done to eliminates unnecessary words and stop words and also to reduce the size of the data that need to be examined.

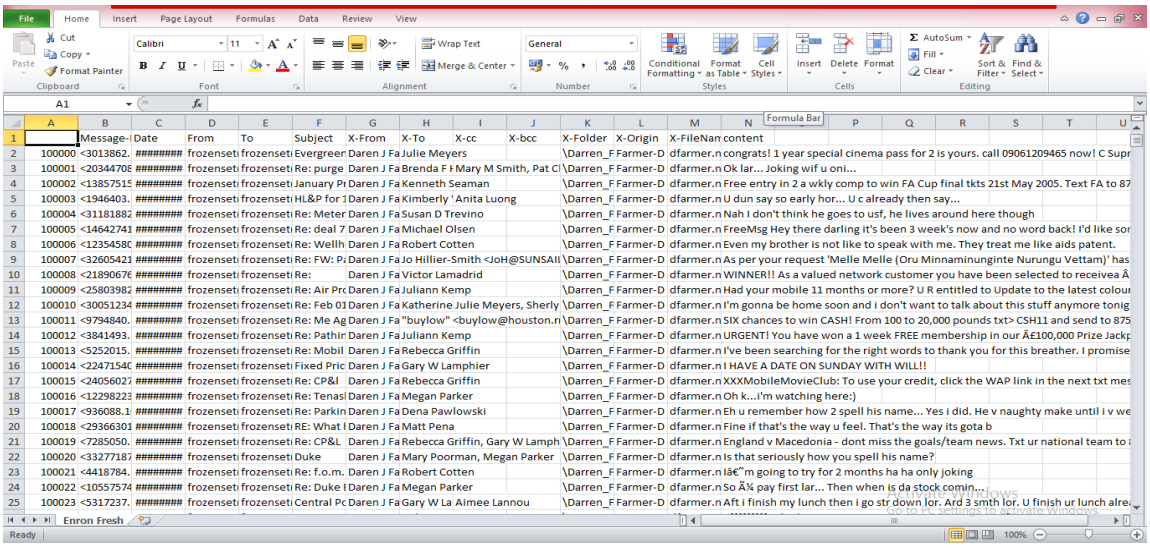

Fig 2: shows data preprocessing of the data set.

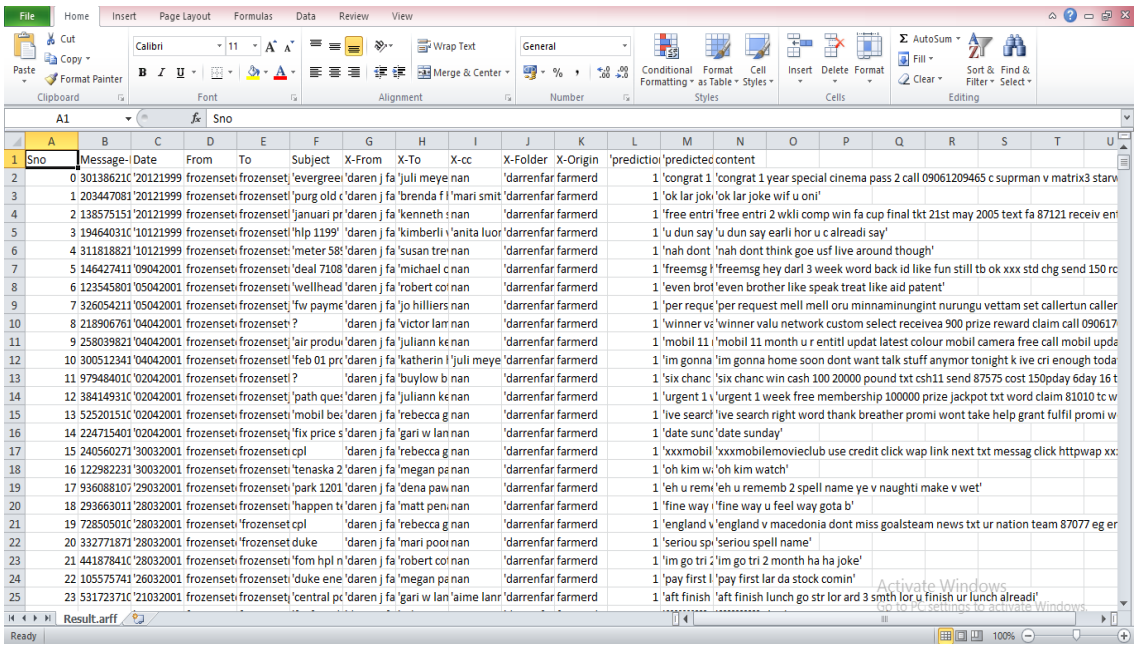

Fig 3: shows feature selection after data preprocessing.

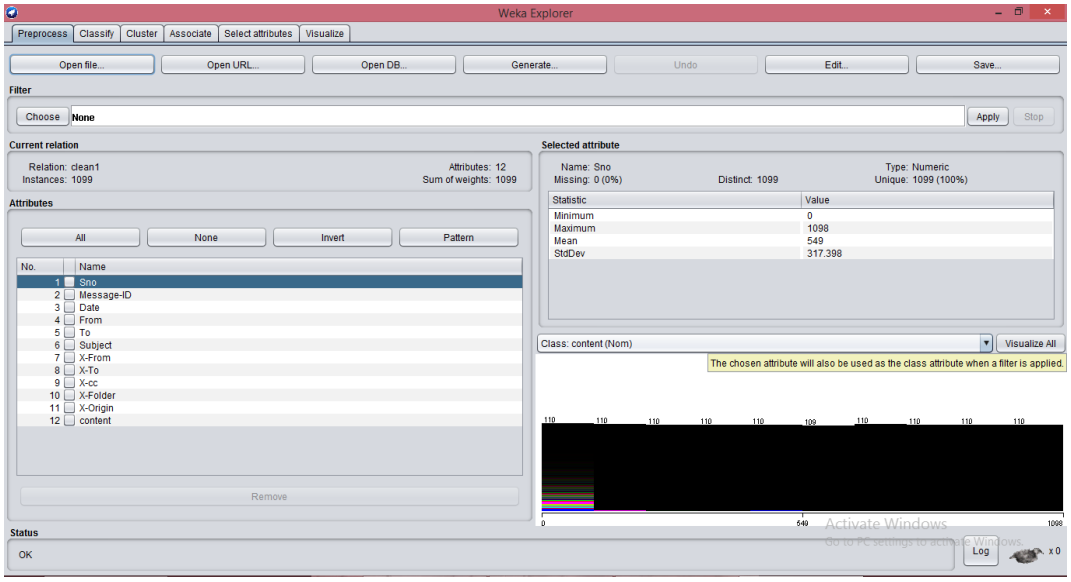

Fig 4: Classification result using weka tool.

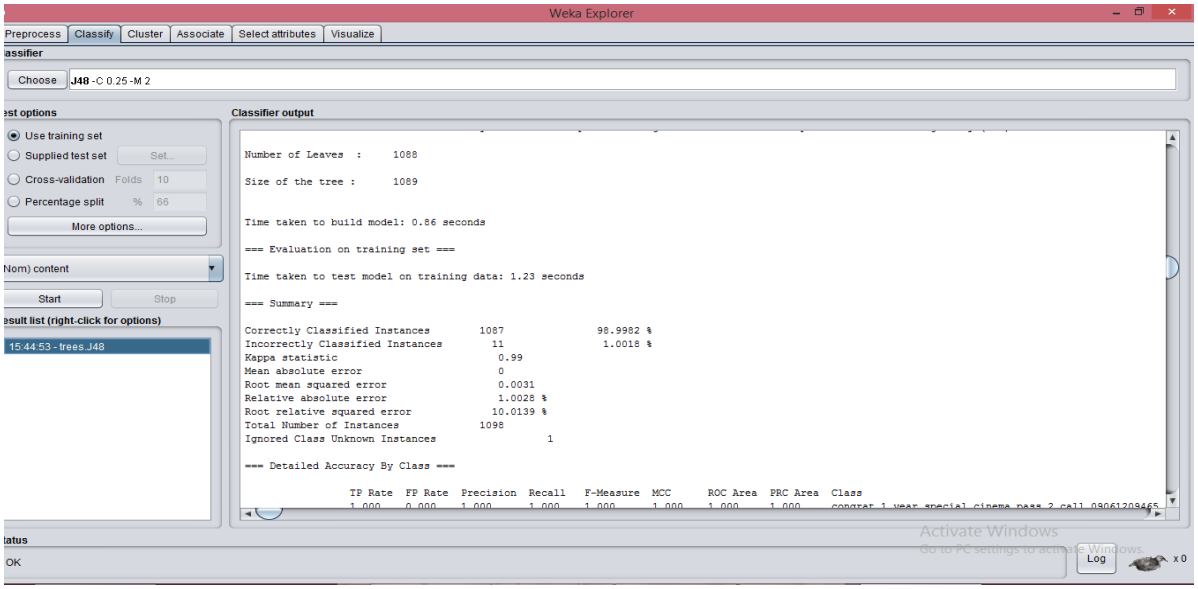

Fig 5: shows classification of E-mails using weka tool.

As shown in Fig 5, this paper uses J48 classifier for classification of Fake E-mails from Genuine E-mails and the result shows that the model could able to classify with 99% accuracy.

VI. CONCLUSION

Even though there are so many methods exist for communication but still E-mails are used widely for official communication, business communication and peer to peer communication. The importance of E-mail communication is growing exponentially day by day. At the same time fraudsters are frequently used E-mails to carry out different attacks. The most important attack is phishing attack uses E-mails as the target to acquire very sensitive information such user name, VISA Card details and passwords. So many researchers have given their ideas for classification of fake E-mails and Real E-mails. Each one having their own limitation, this research work aims to classify between fake E-mails and Genuine E-mails using J48 classification algorithm. It was observed that the classifier able to classify with 98% accuracy, which far better than other research works in this area. Hence the classifier used in this research work is very efficient in terms of accuracy of classification and able to help the user in identifying the Fake E-mails. This research is a helping hand for the common man in protecting their vital data by proper detection of phishing attacks in their E-mails.

REFERENCES

1. P. Liu and T. S. Moh, "Content Based Spam E-mail Filtering," 2016 International Conference on Collaboration Technologies and Systems (CTS), Orlando, FL, pp. 218-224, 2016.
2. N. Agrawal and S. Singh, "Origin (dynamic blacklisting) based spammer detection and spam mail filtering approach," 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), Moscow, pp. 99-104, 2016.
3. Prasanta Kumar Sahoo, “Data Mining a way to Solve Phishing Attacks”, IEEE International Conference on Current Trends towards Converging Technologies (ICCTCT-2018) , Coimbatore, India,2018.

4. Kaspersky Lab, “Spam in January 2012 love, politics and sport,” 2013 ,http://www.kaspersky.com/about/news/spam/2012/Spam_in_January_2012_Love_Politics_and_Sport, 2012.
5. M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabatah, “Modelling intelligent phishing detection system for e-banking using fuzzy data mining,” in CyberWorlds, 2009. CW’09. International Conference on. IEEE, ,pp.265–272, 2009.
6. P. Barraclough, M. Hossain, M. Tahir, G. Sexton, and N. Aslam, “Intelligent phishing detection and protection scheme for online transactions”, Expert Systems with Applications, 2013.
7. A. Almomani, B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, “A survey of phishing E-mail filtering techniques”, Communications Surveys & Tutorials, IEEE, vol. 15, no. 4, pp. 2070–2090, 2013.
8. M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, “Intelligent phishing detection system for e-banking using fuzzy data mining”, Expert systems with applications, vol. 37, no. 12, pp. 7913–7921, 2010.
9. Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F., “Predicting Phishing Websites Using Classification Mining Techniques with Experimental Case Studies”, 2010 Seventh International Conference on Information Technology: New Generations.doi:10.1109/itng.2010.117 , 2010.
10. M. A. Hossain Dept. of Computing University of Brad fordBradford, “Predicting Phishing Websites Using Classification Mining Techniques with Experimental”, UK,2010.
11. https://smallbiztrends.com/2017/08/phishing-examples-small-business.html.
12. http://www.itu.int/ITU-D/e-strategies/e-legislation/Doc/Cybercrime_M_Menting.pdf
13. Choon Lin Tan, Kang Leng Chiew, San Nah Sze , “Phishing Webpage Detection Using Weighted URL Tokens for Identity Keywords Retrieval”, in the proceedings of 9th International Conference on Robotic, Vision, Signal Processing and Power Applications, pp 133-139, Springer Singapore, 2017.
14. Zhijun Yan, Su Liu, Tianmei Wang, Baowen Sun, Hansi Jiang, Hangzhou Yang, “A Genetic Algorithm Based Model for Chinese Phishing E-commerce Websites Detection in HCI in Business”, Government, and Organizations: eCommerce and Innovation, Springer International Publishing, 2016.
15. Yuancheng Li, Rui Xiao, Jingang Feng, Liujun Zhao, “A semi-supervised learning approach for detection of phishing webpages”, Optik-International Journal for Light and Electron Optics, vol.124, Issue 23, December 2013.
16. P. Liu and T. S. Moh, "Content Based Spam E-mail Filtering", 2016International Conference on Collaboration Technologies and Systems(CTS), Orlando, FL, pp. 218-224, 2016.
17. J. Thomas, N. S. Raj and P. Vinod, "Towards filtering spam mails using dimensionality reduction methods," 2014 5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence)” ,Noida, pp. 163-168, 2014.
18. I. R. A. Hamid and J. Abawajy, “Hybrid feature selection for phishing E-mail detection”, in Algorithms and Architectures for Parallel Processing, Springer, pp. 266–275, 2011.

AUTHORS PROFILE

Dr. Prasanta Kumar Sahoo Professor, Department of Computer Science & Engineering, Sreenidhi Institute of Science & Technology, Hyderabad. He completed his Ph.D. from Fakir Mohan University, Odisha in Computer Science Engineering. He has 17 years of teaching, research and administrative experience. He has earlier worked as Head of the Dept. in both CSE and IT dept. in various reputed Engineering Colleges. His Research interest includes Cyber Security, Information Security and Data Mining. He has published around 50 research papers in various reputed journals both at national and International level. His research papers were cited both at national and international level, so far by 41 citation and 1567 reads as per Google Scholar and research Gate report. Many times Dr. Prasanta Kumar Sahoo won the best teacher award in various colleges for his contribution to the teaching and learning process. He is Certified Professional from BalaBit, completed Electronic Contextual Security Intelligence exam Intermediate Level (ECSI). He has guided more than 50 projects both at UG and PG level. He has delivered more than 15 guest lecturers. He has organized three national conference and nine faculty development program with an immense success.

Cheguri Rajitha pursuing Masters Degree in computer Science & Engineering from Sreenidhi Institute of Science and Technology, Hyderabad. She has completed her B. Tech Degree from JNTU Hyderabad. She also published few papers in reputed journals.