International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 11 | Nov 2019 www.irjet.net p-ISSN: 2395-0072

A SURVEY ON AUTOMATIC PHISHING EMAIL DETECTION USING NATURAL LANGUAGE PROCESSING TECHNIQUES

Anirban Mukherjee[1], Nimit Agarwal[2], Shubham Gupta[3]

[1, 2, 3]Vellore Institute of Technology, Vellore, India

Abstract - In today's global age, as the world is getting smaller, making everyone connected to one another; e-mails are an essential tool in accomplishing the interconnection of machines and people across the globe. With the advent of communication technologies and the rise of internet usage in exponential terms over the recent years, e-mails are now very much a daily part of our lives. That being said, it has also unfortunately given rise to many online scams that use phishing techniques to trap unsuspecting users into giving away their personal information that can lead to significant theft of cash, identity theft, character assassination and many other malicious activities having severe consequences for the citizens of the internet. These issues make tacking the challenge of phishing e-mails a priority. Therefore, in this paper, we look into detecting phishing e-mails using Natural Language Processing (NLP) techniques. We delve into the methods that help identify these malicious e-mails using NLP, which helps in detecting and classifying the potentially harmful e-mails so that it does not cause any potential harm to the precious user data. Here, we see how these methods are implemented to detect the e-mails and the work being done to improve upon the current and possible future scenarios because as the scammers keep evolving and get more sophisticated, we need to develop and rise up to the challenge as well.

Key Words - Online Phishing, E-Mail, NLP, Neural Network, Feature Extraction, Spam

1. INTRODUCTION

Phishing is a dishonest endeavor to acquire sensitive and crucial information such as a person's usernames, passwords, credit card details, and effectively, money, for ill-disposed reasons, by masquerading as a reliable entity via electronic communications such as an e-mail. Many users unwittingly click phishing domains in their e-mails every day and every hour. The attackers target both the users and the companies through such e-mails. The main reason for phishing fraud is the lack of awareness of users. Not being able to prevent these mean a huge loss of resources such as data, sensitive information, and of course, money. To prevent this loss of information, one can start with making people conscious and aware. However, the need of the hour is building strong security mechanisms which can detect and prevent phishing domains from reaching the user. Phishers use many different techniques to initiate phishing attacks; the main method primarily being via e-mails. Phishers, however, can easily modify their methods and scamming techniques to use any method available to reach their victims. The phishing attackers trick users by employing different social engineering tactics such as threatening to suspend user accounts if they do not complete the account update process, provide other information to validate their accounts or some other reasons to get the users to visit their spoofed web pages, where they can capture the unsuspecting users' precious data. This can cause identity theft, money stolen using credit card details and can have social repercussions like postage of offensive social media content. Phishers also evolve with time to employ more sophisticated techniques to appear more legitimate and unsuspecting. One such new attack methodology, known as spear phishing, has been utilized often recently. Spear phishing differs from ordinary attacks in subtle ways such as spoofed e-mails coming from somebody the receivers know. As an example, the mail may look like an ordinary mail to employees from their employers asking them to update their data, login information for example, and in the process, capture the data. As the majority of phishing e-mails are formatted to appear from a legitimate source, a large percentage of e-mail users are unable to recognize phishing attacks. Moreover, traditional spam e-mail filters are inclined to fail to identify phishing e-mails since most phishing attacks use more sophisticated techniques and tend to be directed to a more targeted audience. The menace of phishing has been found much more damaging than expected. With the increasing severity of this issue, many efforts have been devoted to applying NLP methods to detect phishing. According to the Anti Phishing Working Group, it was found that there were 18,480 unique phishing attacks and 9666 unique phishing sites reported in March 2006, which has only exponentially increased ever since as per researches. The information security organization Vasco found 2.5 million unique phishing attacks in 2016 alone, showing an almost 100% increase from 2015. Another organization Webroot estimates there are 1.4 million phishing sites created per month. Phishing attacks affect millions of internet users and are a huge cost burden for businesses. The Federal Bureau of Investigation (FBI) research on Business E-Mail Compromise in 2018 showed

that compromising of e-mails due to phishing has cost businesses $12 billion across the world. Thus we see that phishing has become a significant threat to users and businesses alike. Phishing seems to affect e-commerce too, as online customers start doubting the services, and the trust upon the web environment lessens. The most widely known e-mail fraud which victims fall for regardless of the technology used is the 419 Scam. It is one of the longest running scams in the world and seems to be relatively omnipresent. The format is something that seems innocuous and straightforward but is actually scamming the receiver and phishing for passwords and other personal details or even directly asking for money. The scams are along the lines of bogus business, humanitarian or philanthropic related deals where the victim is sent mails where he is promised large sums of money for no initial investment. Attempts to perpetrate this particular type of fraud are regularly experienced by many members of society, including accountants and their clients. The scam originated in Nigeria and is derived from Section 419 of the Nigerian Penal Code that deals with pretenses. Much legislation has been passed, especially in Nigeria, to curb 419 frauds, but without success. Media education campaigns have encouraged the adoption of procedures (including recourse to the advice of accountants) to protect proprietary information and to lead to a safer and more disciplined use of computers and the Internet. Hence we can now see that to be able to detect fraudulent e-mails and caution the user about the same becomes necessary in today's world, and as phishers continue to improve their tactics using new techniques, targeting specific groups, and using alternative channels to spread their attacks. The present research focuses on detecting phishing attacks at the e-mail level since phishers primarily use this channel to initiate their attacks. This paper thus aims to look into the models that utilize NLP techniques for automatic phishing e-mail detection, where it looks into the work that has been done on this aspect and how we are evolving to tackle the ever-increasing challenges presented through such scams.

2. BACKGROUND STUDY

We can now establish that anti-phishing exploration is one of the most import ongoing research areas and fields in the area of data security. Due to the unavailability of an openly standard test dataset, the significant portions of the specialists are utilizing their very own dataset for analysis. This makes the analysis and benchmarking a lot different over various anti-phishing methods to become testing and wasteful. The discoveries of this prerequisite examination have finished up a few impacting factors that will upgrade the dataset quality, which incorporates: the sort of crude components, the wellspring of the example, test size, site class, classification conveyance, the language of the site and the help for highlight extraction. This paper has brought up the interest of the disconnected dataset regarding the anti-phishing research network and has developed a disconnected dataset using NLP. The commitments of this paper include: giving a disconnected dataset to quick primer technique testing and filling in as a changeless archive for phishing pages. Since phishing sites have a short life expectancy, getting to a shutdown phishing site to recover some one of a kind phishing trademark winds up conceivable through this dataset. [1]

We have seen how phishing e-mails are one of the most significant dangers on the planet today and has caused colossal financial losses. Although the strategies for such encounters are continually being updated, the consequences of those techniques are not exceptionally palatable at present. Also, phishing messages are developing at an alarming rate as of late. To tackle this, the authors propose an e-mail phishing detection model called “THEMIS.” To assess the adequacy of THEMIS, the authors have utilized a lopsided dataset that has reasonable count and proportions of phishing and genuine messages. In this paper, messages are separated into two classifications, authentic messages and phishing messages. Typically, the recognition for phishing messages is additionally a parallel classification issue. The authors mathematize the issue and split an e-mail into two sections, the header, and the body. In this paper, the authors have utilized this profound learning model named THEMIS to recognize phishing messages. The model uses an improved Recurrent Convolutional Neural Networks (RCNN) to display the e-mail header and the e-mail body at both the character level and the word level. In this manner, the commotion is brought into the model insignificantly. In the model, the authors have utilized the consideration system in the header and the body, making the model give more consideration to the more significant data between them. The lopsided dataset closer is utilized to this present reality circumstance to direct examinations and assess the model. In the experiments, the model has shown a lot of promise as the accuracy rate of THEMIS has been an impressive 99.848%, with research being done for detecting phishing e-mails using only the e-mail body. [2]

In this paper, the authors propose an innovative new model based on a linear kernel Support Vector Machine (SVM) utilizing robust NLP techniques. The proposed model comprises of two stages: high-level feature extraction and machine learning part. Since there is no present method to have a machine sort messages without removing highlights, there must be an element extraction segment. This model right now utilizes 26 highlights to decide if messages are phishing or ham. While a portion of the highlights may not appear to be powerful, even highlights, for example, word check, can be genuinely successful on its own. Since diverse machine learning models capacity best in specific sets of

information, 17 models were tried. Out of the 17 tried, 14 gave satisfactory outcomes. All models and measurements will be expressly named as weighted or unweighted, and this is the main weighting plan utilized in the whole paper. The authors made phishing e-mail identifiers using novel highlights. The indicator had the option to appropriately distinguish over 80% of phishing messages and 95% of ham messages, while just utilizing 26 highlights. With a significant bit of the highlights concentrating on word checks, stopword tallies, accentuation checks, and uniqueness, the vast majority of the highlights are along these lines novel. [3]

It is evident that phishing is utilized to force people into playing out specific activities or disclosing delicate data. The aggressors may give off an impression of being ordinary, valid, dependable individuals. By posing inquiries, they may sort out enough data to access individual records for banking, e-mail, business, or shopping. In this application, each subject is solicited to audit a rundown from 100 instant messages that seem to originate from a known contact on their phone and afterward distinguish suspicious instant messages. Of these 100 messages, 80 are true, while 20 are imitations. (Subjects were not told what numbers of messages were phony.) Of the 20 falsifications, 10 are conventionally worded, standard phishing messages. The other 10 messages were made utilizing the Stanford Core Natural Language Processor. This examination in-progress is as yet in progress. The after-effects of the pilot study will 0be displayed at the meeting. The outcomes will be scored regarding false positives (inaccurately ordering an authentic message as a phishing endeavor) and false negatives (mistakenly characterizing a phishing endeavor as a real message). [4]

We have seen how phishing attacks have turned out to be one of the most customarily utilized social building techniques in day by day life. Since the assailant doesn't depend on specific vulnerabilities, social building, particularly phishing assaults, can't be handled utilizing digital security instruments like firewalls, IDSs (Intrusion Detection Systems), and so forth. The authors propose a powerful model for shielding delicate data from social building assaults. They name this model Security Training and Processing Evaluation (STPE). The model is fundamentally a cycle with five phases:- course training, operation monitoring, text collecting, model examining, score updating. In this paper, the authors planned a structure that estimates the conduct of the social designers and a far-reaching model for depicting mindfulness, estimation, and guard of social building based assaults. They have attempted to propose a half and half multilayered model utilizing natural language processing procedures for protecting the social building based attacks. The model empowers the speedy discovery of a potential assailant attempting to control the unfortunate casualty for uncovering secret data. [5]

Till now, we have seen multiple times how phishing assaults using e-mails are so prevalent today, yet they are among the least secure and guarded systems. Therefore, the authors propose a methodology that utilizes Natural Language Processing procedures to distinguish explanations, which are a sign of phishing attacks. After experiencing the vast majority of the exploration papers from 2014 to 2018 on the theme e-mail Phishing recognition, we can construe that, for the most part, the datasets that are utilized are 'Spamassassin' and Phishing Corpus and these are broadly publicly released dataset and effectively accessible. The Machine Learning strategies generally used to date are SVM, Random Forest, Naïve Bayes, Logistic Regression, Clustering, and the most recent research papers have utilized Machine Learning based systems, for example, Neural networks and LSTM. Some previous works using these have experienced a high accuracy; however, on a little arrangement of data. The authors, therefore, propose a framework that accepts contribution as standard dataset or blend of datasets with considerable measure of information, for instance, Spamassassin or phish corpus dataset that are accessible as an open-source information or we can make self-possessed dataset by collecting the e-mails, and apply machine learning and profound learning procedures, for example, LSTM, Gullible Bayes, SVM, Random Forest and so forth and then concentrate highlights from them, for example, Tag, URL and a lot more which are sustained to the models in order to characterize the e-mail as phished or real. In the current frameworks, a little example dataset is taken, and thus high accuracy is experienced. Right off the bat, utilizing strategies such as clustering, for deciding the conduct of the resultant bunches, ends up being troublesome. [6]

In this paper, the authors propose a solution where URL locations are utilized in accentuation assaults were attempted to be recognized by using NLP (Natural Language Processing) methods. A phishing location framework that can identify this sort of assault by utilizing some particular calculations utilized in AI and distinguishing some visual similitude with the assistance of some regular language preparing strategies. An assailant who wishes to play out a phishing assault uses the accompanying fundamental techniques to expand assault execution and take more client data: typosquatting, cybersquatting, joined word utilization, and utilization of arbitrary characters. Different modules are made to dissect every technique for phony. Paper demonstrates the execution of Word Decompose Module (WDM), Random Word Detection Module (RWDM), and Maliciousness Analysis Module (MAM). This paper further investigations the after-effects of the three calculations: Random Forest (RB), which is a tree-based calculation, Sequential Minimal Optimization (SMO), which is a portion based calculation and Naive Bayes algorithm (NB). The

Random Forest Algorithm was seen to be more effective than the other algorithms tried with a 97.2% achievement rate. [7]

To prevent phishing, here, the authors employ a supervised system where the creator gives a Hierarchical Long Short Term Memory systems (H-LSTMs) and consideration components to show the messages at the same time at the word and the sentence level. The significant advantage of profound learning is its capacity to naturally initiate compelling and task - explicit portrayals from information that can be utilized as highlights to perceive phishing messages. In the various levelled LSTM model messages are considered as progressive structures with words in the lower level (the word level) and sentences in the upper level (the sentence level). The yields of the LSTM models in the two levels are joined utilizing the consideration instrument that doles out commitment loads to the words and sentences in the messages. The paper proposes the procedure to manage the consideration instrument at the word level of the progressive LSTMs dependent on the appearance rank of the words in the vocabulary. It utilizes the precision, recall, and F1-score to assess the exhibition of the models for identifying phishing messages. Moreover, the proposed models H-LSTMs and H-LSTMs with the supervised approach use the header arrange in the assessment on information full-header. [8]

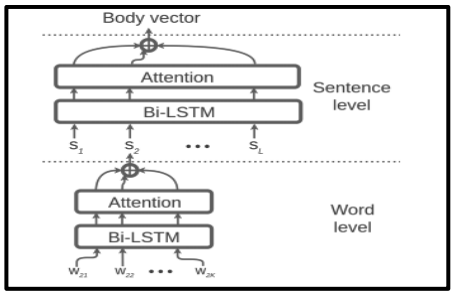

Fig -1: Hierarchical LSTMs

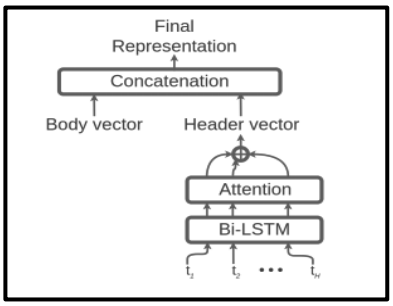

Fig -2: Hierarchical LSTMs with header network

To identify phishing features without any human input. The authors propose this novel study into the phishing content classifier based on a Recurrent Neural Network (RNN). The design and implementation of the paper include the following steps: Firstly, the data has to be pre-processed. This pre-processed data is then fed into the binary RNN classifier. It is customary to 'feed' RNNs with an n-gram representation of the abstracted text. Then the text of the e-mail is tokenized. The naïve approach of splitting on white space characters does not generalize well to e-mail tokenizing. So the paper uses other way of splitting by using special characters and allowing only restricted words in the URL. Finally, the tokenized words are classified by RNN using the logistic sigmoid function. The experimental results of the paper suggest that the methodology has the edge over the current systems, and the proposed system can be enhanced by changing and introducing new methods in the future. [9]

The authors utilize a model utilizing word inserting or vectorization, and they propose a neural system based model for discovery and arrangement of phishing e-mails. The model includes six parts and uses six highlights and ten times cross approval for preparing, approval, and testing. The info highlights are separated from two publically accessible e-mail datasets for both benevolent and phishing messages. The proposed execution utilizes six distinct modules in the plan of the framework – E-mails, E-mail Classifier, E-mail Parser, E-mail Sanitizer, E-mail Vectorizer, and Neural Network Model. The primary segment contains defame and generous messages. The classifier module characterizes arranging each e-mail as either a considerate e-mail or a phishing e-mail. The e-mail parser segment is liable for discovering the number of connections in the e-mail and any Java Script in the e-mail. The NN module incorporates x information sources and y yields associated with coordinated shrouded layers. The neural system works in three layers - preparing layer, approval layer, and testing layer. The outcomes were caught, spoken to, and broke down in three gatherings of perplexity grid, ROC, and system execution. Tending to the caught outcomes, the paper shows a palatable implementation as far as precision, genuine positive rate, and false-positive rate. [10]

Utilizing Machine Learning and NLP approaches. This model performs a semantic analysis of the text transmitted by the attacker to verify the appropriateness of each sentence. A sentence is considered to be malicious if it inquires sensitive information or commands a performance of an action that might expose personal information. Natural language processing (NLP) techniques are applied to parse each sentence and identify the semantic roles of essential words in the sentence concerning the predicate. The algorithm used here analyses sentences in the e-mail one at a time and finds out if the e-mail has any social engineering attacks. The methodology also uses link analysis to verify URLs in the e-mail. After the analysis, if the found URLs are in the predefined blacklist of URLs, then the e-mail is considered as phishing e-mail. The method ignores the e-mails which only

have images. The results of the classifier are evaluated using the F - score. [11]

In this work, the authors look at simultaneousness on the definitive features which should be used in phishing detection. The paper shows the ways to deal with apply Fuzzy Rough Set (FRS) theory as a device to pick best features from three benchmarked datasets. The selected features are sustained into three routinely used classifiers for phishing distinguishing proof. To evaluate the FRS incorporate assurance in the structure up a generalizable phishing area, the classifiers are set up by an alternate out-of-test enlightening record of 14,000 site tests to keep up a good critical way from classification overfitting. Assessments with other part decision techniques utilized by the related work show the outperformance of FRS in picking efficient features. The most extraordinary F-measure grabbed by the features selected by FRS is 95% using Random Forest classification. Also, there are nine comprehensive features picked by FRS considering the three instructive records. The F-measure worth using this comprehensive rundown of capacities is generally 93% which is an indistinguishable result rather than the FRS execution. [12]

3. CONCLUSIONS

Phishing is one of the most severe cyber-crime threats in our times that reduces customer trust in e-commerce and inflicts loss of billions of dollars to businesses. It also causes abominable social problems such as identity thefts, leading to misuse and misrepresentation of identity and is known to compromise social media accounts. In this paper, we identified e-mails as the primary channel of phishing attacks. E-mails are now a crucial part of our daily lives, and therefore, to prevent malicious attempts at phishing, we look into various models and applications which utilize NLP techniques that successfully detect phishing e-mails. Most phishing e-mails are social engineering schemes that threaten, provoke, or tempt users into giving their valuable information. Using NLP, we can detect the various contexts and contents of a standard scam mail, which helps in identifying them easily. As the phishing scams keep increasing along with scammers employing new strategies and techniques, we utilize various machine learning techniques such as neural networks and deep learning that use NLP which allows us to build different types of models. The models are then trained to detect phishing e-mails and these models can adapt by modifications upon them and as per changes in training and test data that is influenced by the scammers' newer types of attempts. That being said, research on e-mail data to prevent phishing is far from being completed. In the future, work on the exploration of many more under-explored characteristics of e-mail data, mainly focusing on meta-information including timestamp, subject, sender, etc., and incorporating well-refined features to feature transformation will improve the classification results. It is assumed that grouping e-mails based on threads detected before conducting analysis will improve efficiency and accuracy. Thus, we see that much research remains to be done, and the onus remains on us to improve with time as the scammers will try to adapt to changes as well, making this a dynamic research area where many solutions and scenarios remained unexplored, and there remain things yet to be worked on.

REFERENCES

[1] Chiew, K.L., Chang, E.H., Tan, C.L., Abdullah, J. and Yong, K.S.C., 2018. Building standard offline anti-phishing dataset for benchmarking. International Journal of Engineering & Technology, 7(4.31), pp.7-14.

[2] Fang, Y., Zhang, C., Huang, C., Liu, L. and Yang, Y., 2019. Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism. IEEE Access, 7, pp.56329-56340.

[3] Egozi, G. and Verma, R., 2018, November. Phishing Email Detection Using Robust NLP Techniques. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 7-12). IEEE.

[4] Shropshire, J., 2018, December. Natural Language Processing as a Weapon. In Proceedings of the 13th Pre-ICIS Workshop on Information Security and Privacy (Vol. 1).

[5] Thakur, K., Shan, J. and Pathan, A.S.K., 2018. Innovations of Phishing Defense: The Mechanism, Measurement and Defense Strategies. International Journal of Communication Networks and Information Security, 10(1), pp.19-27.

[6] Zalavadia, F., Nevrekar, A., Pachpande, P., Pandey, S. and Govilkar, S., Detecting Phishing Attacks Using Natural Language Processing and Deep Learning Models.

[7] Buber, E., Diri, B. and Sahingoz, O.K., 2017, December. NLP based phishing attack detection from URLs. In International Conference on Intelligent Systems Design and Applications (pp. 608-618). Springer, Cham.

[8] Nguyen, M., Nguyen, T. and Nguyen, T.H., 2018. A deep learning model with hierarchical lstms and supervised attention for anti-phishing. arXiv preprint arXiv:1805.01554.

[9] Halgas, L., Agrafiotis, I. and Nurse, J.R., 2019. Catching the Phish: Detecting Phishing Attacks using Recurrent Neural Networks (RNNs). arXiv preprint arXiv:1908.03640.

[10] Moradpoor, N., Clavie, B. and Buchanan, B., 2017, July. Employing machine learning techniques for detection and classification of phishing emails. In 2017 Computing Conference (pp. 149-156). IEEE.

[11] Peng, T., Harris, I. and Sawa, Y., 2018, January. Detecting phishing attacks using natural language processing and machine learning. In 2018 IEEE 12th International Conference on Semantic Computing (ICSC) (pp. 300-301). IEEE.

[12] Zabihimayvan, M. and Doran, D., 2019. Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection. arXiv preprint arXiv:1903.05675.