International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 05 | May-2021 ISSN: 2582-3930

PHISHING EMAIL DETECTION USING LONG SHORT TERM MEMORY [LSTM]

INDHU K[1], METHALAKSHMI G[2], PRATHIKSHA A[2], PRIYANKA S[2], ANITHA M[3]

1UG Scholar, Department of CSE, Kingston Engineering College, Vellore-59
2UG Scholar, Department of CSE, Kingston Engineering College, Vellore-59
3Asst.Professor, Department of CSE, Kingston Engineering College, Vellore-59

ABSTRACTThe phishing email is one of the significant threats in the world today and has caused tremendous financial losses.Although the methods of confrontation are continually being updated, the results of those methods are not very satisfactory at present. Moreover, phishing emails are growing at an alarming rate in recent years. Therefore, more effective phishing detection technology is needed to curb the threat of phishing emails. In this paper, we first analyzed the email structure. Then we proposed a new phishing email detection model named, which is used to model emails at the email header, the email body, the character level, and the word level simultaneously. To evaluate the effectiveness of, we use an unbalanced dataset that has realistic ratios of phishing and legitimate emails. Meanwhile, the ensure that the filter can identify phishing emails with high probability and filter out legitimate emails as little as possible. This promising result is superior to the existing detection methods and verifies the effectiveness of in detectingphishingemails.

Key Words:LSTM,Phishing email,Legitimate email,attention mechanism.

1.INTRODUCTION

The rapid development of Internet technologies has immensely changed on-line users' experience, while security issues are also getting more overwhelming. The current situation is that new threats may not only cause severe damage to customers' computers but also aim to steal their money and identity. Among these threats, phishing is a noteworthy one and is a criminal activity that uses social engineering and technology to steal a victim's identity data and account information. According to a report from the Anti-Phishing Working Group (APWG), the number of phishing detections in the first quarter of 2018 increased by 46% compared with the fourth quarter of 2017 . According to the striking data, it is clear that phishing has shown an apparent upward trend in recent years. Similarly, the harm caused by phishing can be imagined as well. For phishing, the most widely used and influential mean is the phishing email. Phishing email refers to an attacker using a fake email to trick the recipient into returning information such as an account password to a designated recipient. Additionally, it may be used to trick recipients into entering special web pages, which are usually disguised as real web pages, such as a bank's web page, to convince users to enter sensitive information such as a credit card or bank card number and password. Although the attack of phishing email seems simple, its harm is immense. In the United States alone, phishing emails are expected to bring a loss of 500 million dollars per year. According to the APWG, the number of phishing emails increased from 68,270 in 2014 to 106,421 in 2015, and the number of different phishing emails reported from January to June 2017 was approximately 100,000. In addition, Gartner's report notes that the number of users who have ever received phishing emails has reached a total of 109 billion . Microsoft analyzes and scans over 470 billion

emails in Office 365 every month to find phishing and malware. From January to December 2018, the proportion of inbound emails that were phishing emails increased by 250%. Great harm and strong growth momentum have forced people to pay attention to phishing emails. Therefore, many detection methods for phishing emails have been proposed Various techniques for detecting phishing emails are mentioned in the literature. In the entire technology development process, there are mainly three types of technical methods including blacklist mechanisms, classification algorithms based on machine learning and based on deep learning. From previous work, the existing detection methods based on the blacklist mechanism mainly rely on people's identification and reporting of phishing links requiring a large amount of manpower and time. However, applying artificial intelligence (AI) to the detection method based on a machine learning classification algorithm requires feature engineering to manually find representative features that are not conducive to the migration of application scenarios. Moreover, the current detection method based on deep learning is limited to word embedding in the content representation of the email. These methods directly transferred natural language processing (NLP) and deep learning technology, ignoring the specificity of phishing email detection so that the results were not ideal.

2. LITERATURE SURVEY

2.1.TITLE :Automatic Rogue Email Spotter
AUTHORS :C.Coyotees,V.S.Mohan&J.Naveen
YEAR :2018
ABSTRACT:Phishing emails have always bothered users as it's a huge waste of storage, time, money and resource to any user.This work uses word embedding as text representation for supervised classification approach to identify phishing emails.Deep learning based models have shown to surpass the older techniques in spam email detection.
This work aims at attempting the same using a CNN/RNN/MLP network with Word2vec embeddings on phishing email corpus, where Word2vec helps to capture the synaptic and semantic similarity of phishing and legitimate emails in an email corpus.

2.2TITLE:Phishing Email Detection using Classical Machine Learning Technique
AUTHORS :A.Vazhiyil,N.B.Hasikrv&R.Vinayakumar.
YEAR :2018
ABSTRACT;Phishing is a common attack on credulous people by making them to disclose their unique information using counterfeit websites. The objective of phishing website URLs is to purloin the personal information like user name, passwords and online banking transactions. Phishers use the websites which are visually and semantically similar to those real websites.

As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing mechanisms to detect phishing. Machine learning is a powerful tool used to strive against phishing attacks. This paper surveys the features used for detection and detection techniques using machine learning

2.3.TITLE :A deep learning model with hierarchical LSTMs and supervised attention for anti-phishing
AUTHORS :M.Nguyen,T.Nguyen,T.H.Nguyen.
YEAR :2018
ABSTRACT: Anti-phishing aims to detect phishing content/documents in a pool of textual data. This is an important problem in cyber security that can help to guard users from fraudulent information. Natural language processing (NLP) offers a natural solution for this problem as it is capable of analyzing the textual content to perform intelligent recognition. In this work, we investigate state-of-the-art techniques for text categorization in NLP to address the problem of anti-phishing for emails (i.e, predicting if an email is phishing or not).

These techniques are based on deep learning models that have attracted much attention from the community recently. In particular, we present a framework with hierarchical long short-term memory networks (H-LSTMs) and attention mechanisms to model the emails simultaneously at the word and the sentence level.

2.4.TITLE:Deep Learning based Phishing email detection
AUTHORS :M.Hiransha,N.A.Unnithan,R.Vinayakumar.
YEAR :2018
ABSTRACT: Email communication, has now become an inevitable communication tool in our daily life. Especially for finance sector, communication through email plays an important role in their businesses. So, it is very important to classify emails based on their behavior. Email phishing one of most dangerous Internet phenomenon that cause various problems to business class mainly to finance sector.

This type of emails steals our valuable information without our permission, more over we won0 t be aware of such an act even if it has been occurred. In this paper, we reveal about how to distinguish phishing emails from legitimate mails. Dataset had two types of email texts one with header and other without header. We used Keras Word Embedding and Convolutional Neural Network to build our model. This work aims to show the abilities of word embedding have to solve issues related to cybersecurity use cases.

2.5.TITLE:Using Syntactic Feature for Phishing Detection
AUTHORS :G.Park & J.M.Taylor
YEAR :2015
ABSTRACT:This paper reports on the comparison of the subject and object of verbs in their usage between phishing emails and legitimate emails.The purpose of this research is to explore whether the syntactic structures and subjects and objects of verbs can be distinguishable features for phishing detection. To achieve the objective, we have conducted two series of experiments: the syntactic similarity for sentences, and the subject and object of verb comparison. The results of the experiments indicated that both features can be used for some verbs, but more work has to be done for others.

2.6.TITLE:Phishing email detection technique by using hybrid feature
AUTHORS :L.M.Form,K.L.Chiew,S.N.Sze,W.K.Tiong.
YEAR :2015
ABSTRACT:Phishing emails is growing at an alarming rate in this few years. It has caused tremendous financial losses to internet users. Phishing techniques getting more advance everyday and this has created great challenge to the existing anti-phishing techniques. Hence, in this paper, we proposed to detect phishing emails through hybrids features. The hybrid features consist of content-based, URL-based, and behavior-based features.

Based on a set of 500 phishing emails and 500 legitimate emails, the proposed method achieved overall accuracy of 97.25% and error rate of 2.75%. This promising result verify the effectiveness of the proposed hybrid features indetectingphishingemail.

2.7.TITLE :Hybrid feature selection for phishing email detection
AUTHORS :I.R.A.Hamid & J.Abawajy
YEAR :2011
ABSTRACT:Phishing emails are more active than ever before and putting the average computer user and organizations at risk of significant data, brand and financial loss. Through an analysis of a number of phishing and ham email collected, this paper focused on fundamental attacker behavior which could be extracted from email header. It also put forward a hybrid feature selection approach based on combination of content-based and behavior-based.

The approach could mine the attacker behavior based on email header. On a publicly available test corpus, our hybrid features selections are able to achieve 96% accuracy rate. In addition, we successfully tested the quality of our proposed behavior-based feature using the information gain.

2.8.TITLE :New filtering approaches for phishing email
AUTHORS :Fraunhofer & K.U.Leuvan
YEAR :2009
ABSTRACT:Phishing emails usually contain a message from a credible looking source requesting a user to click a link to a website where she/he is asked to enter a password or other confidential information. Most phishing emails aim at withdrawing money from financial institutions or getting access to private information. Phishing has increased enormously over the last years and is a serious threat to global security and economy. There are a number of possible countermeasures to phishing.

These range from communication-oriented approaches like authentication protocols over blacklisting to content-based filtering approaches. We argue that the first two approaches are currently not broadly implemented or exhibit deficits. Therefore content-based phishing filters are necessary and widely used to increase communication security. A number of features are extracted capturing the content and structuralpropertiesofthemail.

2.9.TITLE :An empirical analysis of phishing blacklist
AUTHORS :S.Sheng,B.Wardman,G.Warner,L.Cranor.
YEAR :2009

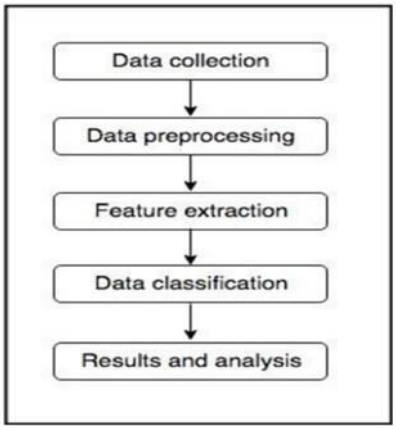
ABSTRACT:In this paper, we study the effectiveness of phishing blacklists. We used 191 fresh phish that were less than 30 minutes old to conduct two tests on eight anti-phishing toolbars. We found that 63% of the phishing campaigns in our dataset lasted less than two hours. Blacklists were ineffective when protecting users initially, as most of them caught less than 20% of phish at hour zero.

We also found that blacklists were updated at different speeds, and varied in coverage, as 47% - 83% of phish appeared on blacklists 12 hours from the initial test. We found that two tools using heuristics to complement blacklists caught significantly more phish initially than those using only blacklists. However, it took a long time for phish detected by heuristics to appear on blacklists. Finally, we tested the toolbars on a set of 15,345 legitimate URLs for false positives, and did not find any instance of mislabeling for either blacklists or heuristics. We present these findings and discuss ways in which anti-phishing tools can be improved

3.PROPOSED SYSTEM

With the emergence of email, the convenience of communication has led to the problem of massive spam, especially phishing attacks through email. Various anti phishing technologies have been proposed to solve the problem of phishing attacks. studied the effectiveness of phishing blacklists. Blacklists mainly include sender blacklists and link blacklists. This detection method extracts the sender's address and link address in the message and checks whether it is in the blacklist to distinguish whether the email is a phishing email. The update of a blacklist is usually reported by users, and whether it is a phishing website or not is manually identified. At present, the two well-known phishing websites are PhishTank and OpenPhish. To some extent, the perfection of the blacklist determines the effectiveness of this method based on the blacklist mechanism for phishing email detection. The current situation is that new threats may not only cause severe damage to customers' computers but also aim to steal their money and identity. Among these threats, phishing is a noteworthy one and is a criminal activity that uses social engineering and technology to steal a victim's identity data and account information. According to a report from the Anti-Phishing Working compared with the fourth quarter of According to the striking data, it is clear that phishing has shown an apparent upward trend in recent years. Similarly, the harm caused by phishing can be imagined as well.

FLOW CHART



3.1.ALGORITHMS

Let's quickly summarize the different algorithms in the R-CNN family (R-CNN, Fast R-CNN, and Faster R-CNN) that we saw in the first article. This will help lay the ground for our implementation part later when we will predict the bounding boxes present in previously unseen images (new data). R-CNN extracts a bunch of regions from the given image using selective search, and then checks if any of these boxes contains an object. We first extract these regions, and for each region, CNN is used to extract specific features. Finally, these features are then used to detect objects. Unfortunately, R-CNN becomes rather slow due to these multiple steps involved in the process. Fast R-CNN, on the other hand, passes the entire image to ConvNet which generates regions of interest (instead of passing the extracted regions from the image). Also, instead of using three different models (as we saw in R-CNN), it uses a single model which extracts features from the regions, classifies them into different classes, and returns the bounding boxes. All these steps are done simultaneously, thus making it execute faster as compared to R-CNN. Fast R-CNN is, however, not fast enough when applied on a large dataset as it also uses selective search for extracting the regions.

3.2 CONCLUSION

we use a new deep learning model named to detect phishing emails. The model employs an improved RCNN to model the email header and the email body at both the character level and the word level. Therefore, the noise is introduced into the model minimally. In the model, we use the attention mechanism in the header and the body, making the model pay more attention to the more valuable information between them. We use the unbalanced dataset closer to the real-world situation to conduct experiments and evaluate the model. The model obtains a promising result. Several experiments are performed to demonstrate the benefits of the proposed model. For future work, we will focus on how to improve our model for detecting phishing emails with no email header and only an email body..

REFERENCE

[1]C. Coyotes, V. S. Mohan, J. Naveen, R. Vinayakumar, and K. P. Soman, ‘‘ARES: Automatic rogue email spotter,’’ in Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal. (IWSPA), A. D. R. Verma, Ed. Tempe, AZ, USA, Mar. 2018.

[2] 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.

[3] M. Nguyen, T. Nguyen, and T. H. Nguyen. (2018). ‘‘A deep learning model with hierarchical LSTMs and supervised attention for anti-phishing.’’ [Online]. Available: https://arxiv.org/abs/1805.01554

[4] M. Hiransha, N. A. Unnithan, R. Vinayakumar, and K. Soman, ‘‘Deep learning based phishing e-mail detection,’’ in Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal. (IWSPA), A. D. R. Verma, Ed. Tempe, AZ, USA, Mar. 2018.

[5] G. Park and J. M. Taylor. (2015). ‘‘Using syntactic features for phishing detection.’’ [Online]. Available: https://arxiv.org/abs/1506.00037

[6] L. M. Form, K. L. Chiew, S. N. Sze, and W. K. Tiong, ‘‘Phishing email detection technique by using hybrid features,’’ in Proc. 9th Int. Conf. IT Asia (CITA), Aug. 2015, pp. 1–5.

[7] R. Verma and N. Hossain, ‘‘Semantic feature selection for text with application to phishing email detection,’’ in Proc. Int. Conf. Inf. Secur. Cryptol. Cham, Switzerland: Springer, 2013, pp. 455–468,

[8] I. R. A. Hamid and J. Abawajy, ‘‘Hybrid feature selection for phishing email detection,’’ in Proc. Int. Conf. Algorithms Archit. Parallel Process. Berlin, Germany: Springer, 2011, pp. 266–275.

[9]Franunhofer,K.U.Leuvan,”New Filtering approaches For Phishing Email”,(2009)

[10] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, ‘‘An empirical analysis of phishing blacklists,’’ in Proc. 6th Conf. Email Anti-Spam (CEAS), Sacramento, CA, USA, 2009, pp. 1–10.

[11] Anti-Phishing Working Group. (2018). Phishing Activity Trends Report 1st Quarter 2018. [Online]. Available: http://docs.apwg.org/Preports/apwg_trends_report_q1_2018.pdf

[12] PhishLabs. (2018). 2018 Phish Trends & Intelligence Report. [Online]. Available: https://info.phishlabs.com/hubfs/2018%20PTI %20Report/PhishLabs%20Trend %20Report_2018-digital.pdf

[13] Anti-Phishing Working Group. (2016). Phishing Activity Trends Report 4th Quarter 2016. [Online]. Available: http://docs.apwg.org/ reports/apwg_trends_report_q4_2016.pdf

[14] Anti-Phishing Working Group. (2015). Phishing Activity Trends Report 1st-3rd Quarter 2015. [Online]. Available:http://docs.apwg.org/ Preports/apwg_trends_report_q1-q3_2015.pdf