International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

An Intelligent Anti Phishing Technology Based On RCNN for Detection Of Phishing Websites

Prof. Guruprasad kulkarni[1], Aishwarya[2], Amrita Roy[3], Aparna Kulkarni[4]

[1,2,3,4]*Dept Of Computer Science and Engineering,*
Guru Nanak Dev Engineering College, Karnataka, India

Abstract- Phishing emails are one of the significant threats in the world today and have caused tremendous financial losses. Although the methods of confrontation are continually being updated, the results of those methods are not very satisfactory at present. Therefore, more effective phishing detection technology is needed to curb the threat of phishing emails. In this project, we first analyzed the email structure. Then, based on an improved recurrent convolutional neural networks (RCNN) model with multilevel vectors and attention mechanism, we proposed a new phishing email detection model named THEMIS, 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 THEMIS, we use an unbalanced dataset that has realistic ratios of phishing and legitimate emails. The experimental results show that the overall accuracy of THEMIS reaches 99.848%. Meanwhile, the false-positive rate (FPR) is 0.043%. High accuracy and low FPR 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 THEMIS in detecting phishing emails.

***Kewords*: Phishing, RCNN model, THEMIS model, Machine learning**

1. INTRODUCTION

Phishing Is one of the major problems faced by the cyber-world and leads to financial losses for both Industries and Individuals. Detection of a phishing attack with high accuracy has always been a challenging issue. At present, visual similarities based techniques are very useful for detecting phishing websites efficiently. The phishing website looks very similar in appearance to its corresponding legitimate website to deceive users into believing that they are browsing the correct website. Visual similarity-based phishing detection techniques utilize the feature set like text content, text format, HTML tags, Cascading Style Sheet (CSS), image, and so forth, to make the decision. These approaches compare the suspicious website with the corresponding legitimate website by using various features and if the similarity is greater than the predefined threshold value then it is declared phishing. Ibis paper presents a comprehensive analysis of phishing attacks, their exploitation, some recent visual similarity-based approaches for phishing detection, and its comparative study. Our survey provides a better understanding of the problem, current solution space, and the scope of future research to deal with phishing attacks efficiently using visual similarity-based approaches.

Composes a bogus e-mail and sends it to the thousands of users. Attacker attached the URL of the fake website in the bogus e-mail. In the case of a spear-phishing attack, an attacker sends the e-mail to selected users.

Nowadays, as there are so many people are being aware of using the internet to perform various activities like online shopping, online bill payment, online mobile recharge, banking transaction. Due to the wide use of this customer face various security threats like cybercrime. Many cybercrimes are widely performed for example spam, fraud, cyber terrorisms, and phishing. Among this phishing is a new cybercrime and very popular nowadays. Phishing is a fraud attempt, which performed to obtain sensitive information about the user. The phisher design website looks the same as any legitimate site and spoof user for obtaining private information of the user such as username, password, banking details for miscellaneous reasons.

2.PROBLEM OVERVIEW

Email communication has now become an inevitable communication tool in our daily life. Email phishing one of the most dangerous Internet phenomenon that causes various problems to business class mainly to the finance sector. This type of email steals our valuable information without our permission, moreover, we won0 t be aware of such an act even if it has occurred. In this paper, we reveal how to distinguish phishing emails from legitimate emails. Dataset had two types of email texts one with header and another without a header. We used Keras Word Embedding and Convolutional Neural Network to build our model.

2.1 PHISHING MECHANISM

The phishing mechanism is shown in Figure. The fake website is the clone of a targeted genuine website, and it always contains some input fields (e.g., text box). When the user submits hasher personal details, the information is transferred to the attacker. An attacker steals the credential of the innocent user by performing the following steps:

Construction of Phishing Site

In the first step, the attacker identifies the target as a well-known organization. Afterward, attacker collects the detailed information about the organization by visiting its website. The attacker then uses this information to construct a fake website. *URL Sending.* In this step, the attacker composes a bogus e-mail and sends it to the thousands of users. Attacker attached the URL of the fake website in the bogus e-mail. In the case

of a spear-phishing attack, an attacker sends the e-mail to selected users.

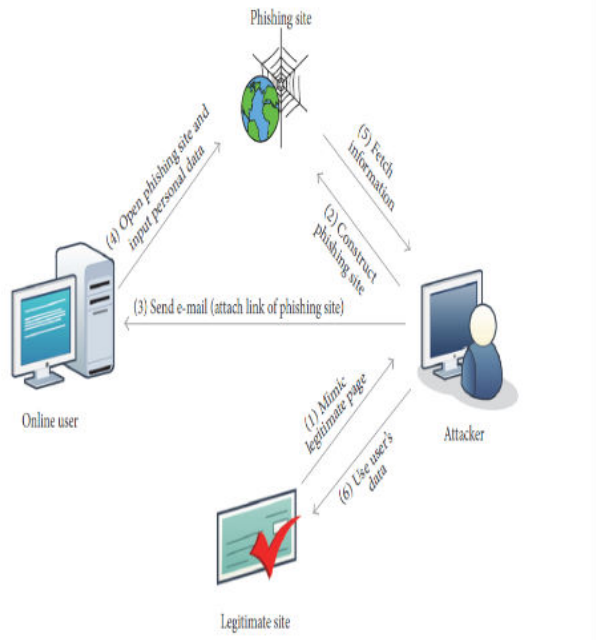


Fig 1: Phishing website construction by attacker

3.RELATED WORK

As in any other system development model, system analysis is the first phase of development in the case of Object Modeling too. In this phase, the developer interacts with the user of the system to find out the user requirements and analyze the system to understand the functioning.

3.1 Existing System

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. 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.

3.2. Proposed System

The proposed detection method is 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 Given the methods mentioned above and the corresponding problems, we set to study phishing email detection systematically based on deep learning.

Concerning the particularity of the email text, we analyze the email structure, and mine the text features from four more detailed parts: the email header, the email body, the word-level, and the char-level.

The RCNN model is improved by using the Then, the email is modeled from multiple levels using an improved RCNN model. Noise is introduced as little as possible, and the context information of the email can be better captured

The THEMIS model proposed in this paper performs well on an unbalanced dataset. The accuracy achieves 99.848%, and all evaluation metrics of THEMIS are superior to the existing detectiontechnologies.

4. SYSTEM DESIGN

4.1 ARCHITECTURE

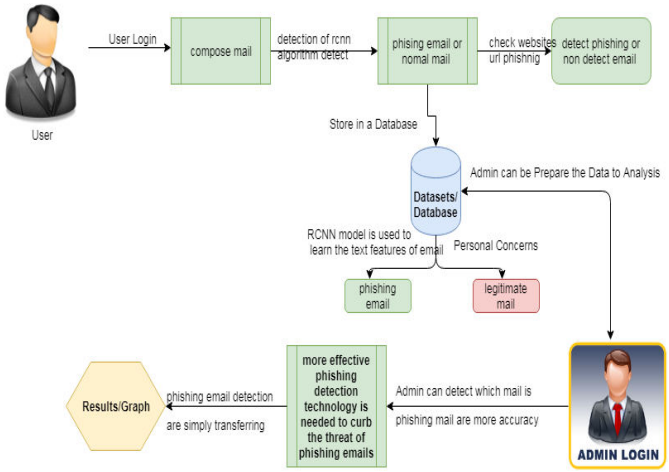


Fig 2: System architecture

4.2 FRAMEWORK

Fig.3 shows our framework for classifying legitimate and phishing emails. First, because the content of the email is very irregular, we need to simply process the email dataset replace and delete the extra spaces and digital gibberish in the text. The email is divided into multiple levels: the char-level and the word-level of the email header as well as the char-level and the word-level of the email body. Then, Word2Vec [40] is used to train and obtain the sequences of vectors. Next, we divide the data into two parts, one as a training-validation set and the other as a testing set. We input a part of the training-validation set into our model and train it to obtain the classifier. Additionally, the other part of the training-validation set is used to carry out the super-parameter selection experiment on the classifier to obtain the beat classification threshold. Finally, the testing set is used to test the determined classifier model to verify the function.

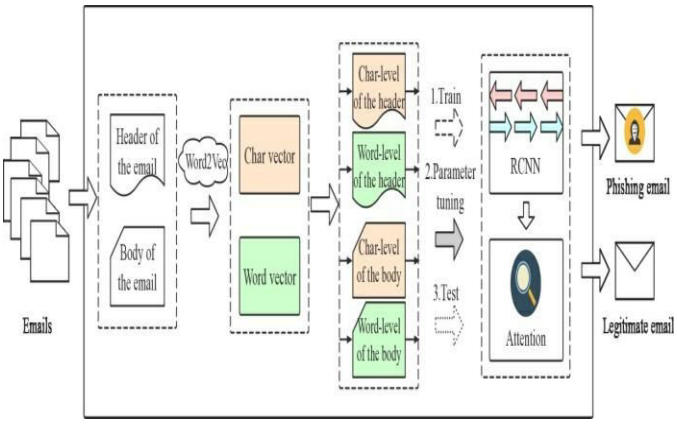


Fig 4: Framework to classify legitimate and phishing mails

4.3 RCNN AND THEMIS MODEL

In deep learning, a **convolutional neural network** (**CNN**, or **ConvNet**) is a class of deep neural networks, most commonly applied to analyzing visual imagery. They are also known as **shift invariant** or **space invariant artificial neural networks** (**SIANN**), based on their shared-weights architecture and translation invariance characteristics. They have applications in image and video recognition, recommender systems,image classification, medical image analysis, natural language processing, and financial time series.

CNN's are regularized versions of a multilayer perceptron. Multilayer perceptron usually means fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer.

A **recurrent neural network** (**RNN**) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. Derived from feedforward neural networks, RNNs can use their internal state (memory) to process variable-length sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition, or speech recognition.

1. Fully recurrent

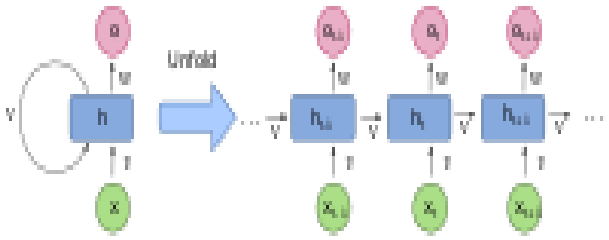


Fig 5: Unfolded basic recurrent neural network

Basic RNNs are a network of neuron-like nodes organized into successive layers. Each node in a given layer is connected with a directed (one-way) connection to every other node in the next successive layer. Each node (neuron) has a time-varying real-valued activation. Each connection (synapse) has a modifiable real-valued weight. Nodes are either input nodes (receiving data from outside the network), output nodes (yielding results), or hidden nodes (that modify the data *en route* from input to output).

THEMIS MODEL

Based on the multilevel embedding and improved RCNN Attention model mentioned earlier, we formally put forward the THEMIS model, a phishing email detection model by combining these two parts. We vectored the email according to its text structure: header, body, characters, and words, using Word2Vec to output char-level embedding and word-level embedding. We combine the text structure of the email into the char-level of the email header, word-level of the email header, char-level of the email body, and word-level of the email body, and we input them into the embedding layer and Bi-LSTM. The result obtained is then connected to the result obtained by the embedding layer to form a triple representation. Then, the triples are sequentially inputted into the dense layer and longitudinal max polling. The results are connected according to the header and the body to obtain the representation of the email header and the representation of the email body. As mentioned in subsection B, different features of email can be mined from the email header and the email body respectively. The email body mainly contains semantic features, and the email header mainly contains routing features and a small part of semantic features. Sometimes, in the body of an email, most phishing emails are nearly identical to the legitimate email. In this case, we should give low weight to the content of the email body to ensure the accuracy of the phishing email detection results.

5. EXPERIMENTAL RESULTS

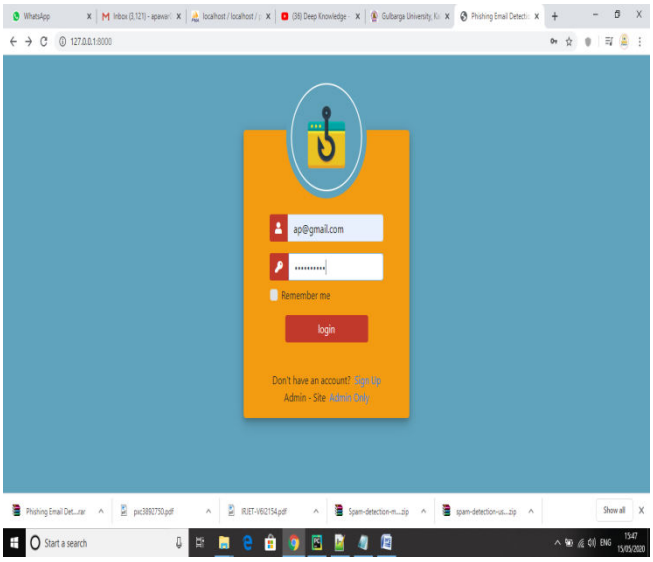
We use regular expressions to match than delete blank lines and unnecessary spaces that still exist in the email body.to the greatest extent, we reduce noise and keep as much email information as possible. python library called email which is a standard parser that understands most email document structures is used to extract the email headers from processed data and then divide email into two parts, namely email header and email body. In the email header part, we save each field as a key-value pair. At the same time, the email header part and email body part are segmented by word segmentation and character segmentation. Email is different from other texts and its content has some uniqueness. In the vectorization phase, it is necessary to train the word vector and character vector for email data.

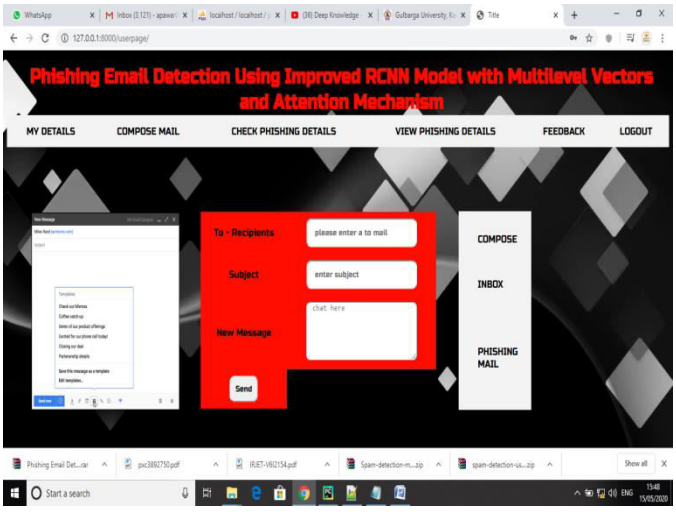
The procedure level testing is made first. By giving improper inputs, the errors occurred are noted and eliminated. Then the web form level testing is made. For example storage of data to the table in the correct manner. In the company as well as the seeker registration form, the zero-length username and password are given and checked. Also, the duplicate username is given and checked. In the job and question entry, the button will send data to the server only if the client-side validations

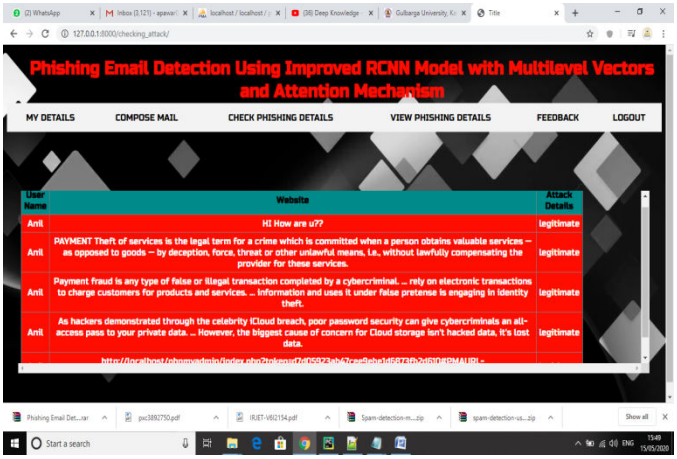
are made. The dates are entered the wrong manner and checked. A wrong email-id and website URL (Universal Resource Locator) is given and checked.

Testing is done for each module. After testing all the modules, the modules are integrated and testing of the final system is done with the test data, specially designed to show that the system will operate successfully in all its aspects conditions

The final step involves Validation testing, which determines whether the software function as the user expected







6.CONCLUSION

In this project, we use a new deep learning model named THEMIS 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 THEMIS model obtains a promising result. Several experiments are performed to demonstrate the benefits of the proposed THEMIS 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.

7.REFERENCES

[1] 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

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

[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] 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

[5] 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

[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] J. Peng, K.-K. R. Choo, H. Ashman, "Astroturfing detection in social media: Using binary n-gram analysis for authorship attribution", *Proc. 15th IEEE Int. Conf. Trust Secure. Privacy Comput. Commun. (TrustCom)*, pp. 121-128, 2016.

[8] 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 Secure. Privacy Anal. (IWSPA), A. D. R. Verma, Ed. Tempe, AZ, USA, Mar. 2018.

[09]. M. Hiransha, N. A. Unnithan, R. Vijayakumar, K. Soman, A. D. R. Verma, "Deep learning-based phishing e-mail detection", *Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secure. Privacy Anal. (IWSPA)*, Mar. 2018.

[10] L. M. Form, K. L. Chiew, S. N. Sze, W. K. Tiong, "Phishing email detection technique by using hybrid features", *Proc. 9th Int. Conf. IT Asia (CITA)*, pp. 1-5, Aug. 2015.