iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IX **Month of publication:** September 2022

DOI: https://doi.org/10.22214/ijraset.2022.46812

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

Phishing Email Detection Using Improved RCNN with Multilevel Vectors and Attention Mechanism

T. Kavitha[1], G. Vaishnavi[2], M. Ramyashree[3], Charanjeet Kaur[4]
[1]*Assistant Professor,* [2,3,4]*UG Students, Department of CSE, Sridevi Women's Engineering College, Hyderabad, T.S, India*

Abstract: ***The Phishing emails are the common threats in the present world which leads to the financial losses and stealing the sensitive information like credit card details, username and password etc. Due to increase in the rate of phishing emails there is a need to introduce the effective phishing detection technology. In order to detect the phishing emails many methods, techniques and technologies are introduced .Improved RCNN with multilevel vectors and attention mechanism is one such technology. Firstly, the email structure is analyzed and then by using RCNN model with multilevel vectors ,attention mechanism phishing content is detected .we introduce noise to test the effectiveness.***

I. INTRODUCTION

Due to the increasing growth rate of internet technologies security has become a major concern for the online users . Emails are oftenly used to exchange data which can be personal or business related matter.The emails can contain sensitive information which phishers want to steal. The Phishers send a email to the receiver which may contain link ,when receiver click on that link and provide information the receivers information might be used for inappropriate purpose .Inorder to detect such phishing emails we are using the methods like RCNN with multilevel vectors and attention mechanism .

The following methods are used to detect the phishing email:

1) First the email structure is analyzed and mine the text features from email header , email body , work-level and char-level.
2) Using RCNN similar patterns from the email is recognized
3) The email goes through many layers for 0the check using RCNN
4) Using attention mechanism different weights are assigned to different parts of the email like email header and email body to focus more on important information .

II. ALGORITHM

A. RCNN Algorithm

RCNN is used for both image and text classification .Text Classification is the the process of categorizing text into organized groups. By using Natural Language Processing (NLP), text classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content .Object detection is the process of finding and classifying objects in an image. One deep learning approach, regions with convolutional neural networks (R-CNN), combines rectangular region proposals with convolutional neural network features. R-CNN is a two-stage detection algorithm. The first stage identifies a subset of regions in an image that might contain an object. The second stage classifies the object in each region .R-CNN and it's faster variant (Faster R-CNN) have shown superior performance for object detection. You could train the R-CNN model to either detect text and background or to detect license plate images .Once you have the object proposal, you can then run text understanding on that using a CNN or any other ML system.

III. COMPONENTS

A. Multilevel Vector

Manual extraction of features is difficult and time taking ,we can not achive the effective results .Hence we opt for Multilevel vectors ,which are very useul to extract the features from image or text . As in the case of Phishing email detection,the email has two parts namely email header and email body ,Multilevel vector checks the email header at charecter-level and word-level . It checks the email body at charecter-level and word-level .Mostly the phishing content can be found in the email body because the structure of email header is mostly same for all the emails but the email body differs from email to email .The email body is more attractive to get the attention from the victim which differs from legitimate mails .The words which are inappropriate and the words which tells about the fraud or crime is detected .

B. Attention Mechanism

Attention is an increasingly popular mechanism used in a wide range of neural architectures .The attention mechanism is a part of a neural architecture that enables to dynamically highlight relevant features of the input data, which, in NLP, is typically a sequence of textual elements. It can be applied directly to the raw input or to its higher level representation. The core idea behind attention is to compute a weight distribution on the input sequence, assigning higher values to more relevant elements .Attention can be used to compare the input data with a query element based on measures of similarity or significance. It can also autonomously learn what is to be considered relevant, by creating a representation encoding what the important data should be similar to .attention is a technique that mimics cognitive attention. The effect enhances some parts of the input data while diminishing other parts — the motivation being that the network should devote more focus to the small, but important, parts of the data. Learning which part of the data is more important than another depends on the context .

C. Neural Networks

An artificial neural network, or neural network, is a mathe-matical model inspired by biological neural networks. Inmost cases it is an adaptive system that changes its structure during learning . There are many different types of NNs. For the purpose of phishing detection, which is basically a classification problem, we choose multilayer feedforward NN. In a feedforward NN, the connections between neurons do not form a directed cycle. Contrasted with recurrent NNs, which are often used for pattern recognition, feedforward NNs are better at modeling relation-ships between inputs and outputs. In our experiments, we use the most common structure of multilayer feedforwardNN, which consists of one input layer, one hidden layer and one output layer. The number of computational units in the input and output layers corresponds to the numberof inputs and outputs. Different numbers of units in the hidden layer are attempted .

D. Deep Learning

Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain. ―Deep‖ refers to the many layers the neural network accumulates over time, with performance improving as the network gets deeper. Each level of the network processes its input data in a specific way, which then informs the next layer. So the output from one layer becomes the input for the next . The adjective "deep" in deep learning refers to the use of multiple layers in the network. Early work showed that a linear perceptron cannot be a universal classifier, but that a network with a nonpolynomial activation function with one hidden layer of unbounded width . Deep neural networks consist of multiple layers of interconnected nodes, each building upon the previous layer to refine and optimize the prediction or categorization .

E. NLP

Natural Language Processing (NLP) is a subfield of artificial intelligence (AI). It helps machines process and understand the human language so that they can automatically perform repetitive tasks. In natural language processing, human language is separated into fragments so that the grammatical structure of sentences and the meaning of words can be analyzed and understood in context. This helps computers read and understand spoken or written text in the same way as humans. Email filters are one of the most basic and initial applications of NLP online. It started out with spam filters, uncovering certain words or phrases that signal a spam message. But filtering has upgraded, just like early adaptations of NLP. NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to _understand‘ its full meaning, complete with the speaker or writer‘s intent and sentiment . NLP also plays a growing role in enterprise solutions that help streamline business operations, increase employee productivity, and simplify mission-critical business processes.

IV. PROPOSED SYSTEM

In this project , Emails are divided into two categories : legitimate emails and phishing emails . A binary variable y is used to represent an email .If y=1 ,the email is a phishing email and y=0 means that email is legitimate . This phishing email detection model is used to model emails at the email header, the email body, the character level, and the word level simultaneously . The email is modelled from multiple levels using an improved RCNN model .The attention mechanism is applied between the email header and and the email body ,and different weights are assigned to the two parts so that the model can focus on more different and more useful information.

V. RESULTS

The phisher who wants to steal the information compose a mail and send to the victim .By using this application you can check whether the website URL is legitimate or phishing .The URL is tested for the phishing content .If the phishing content is present it will be detected .As shown in the picture .

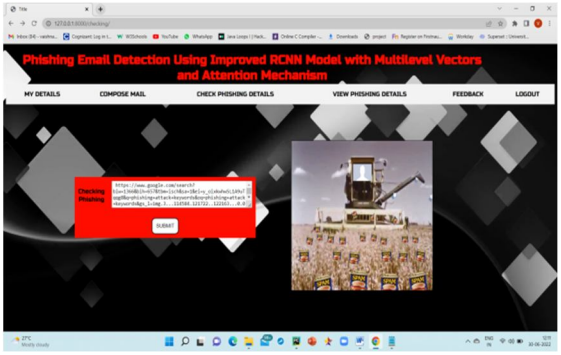


Fig 5.1 Entry of the URL

In the above picture the URL is checked for phishing.

The admin logon to the page and give the analysis of the phishing emails and legitimate emails . Graphical analysis of the phishing and legitimate emails are given .As shown in the given picture

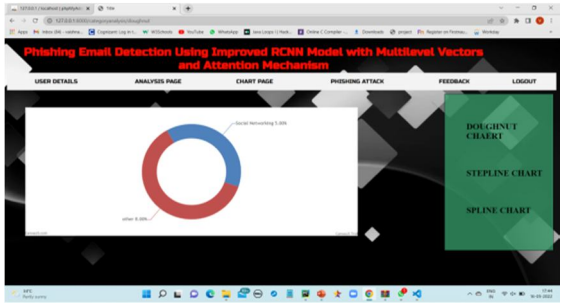


Fig 5.2 Doughtnut Graph

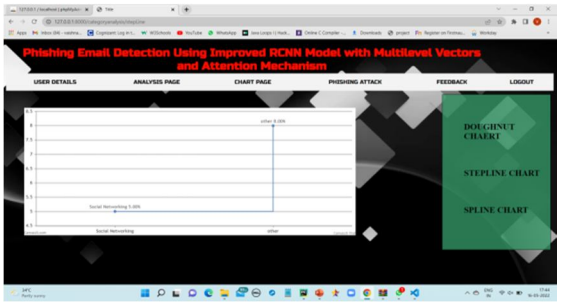


Fig 5.3 Stepline Graph

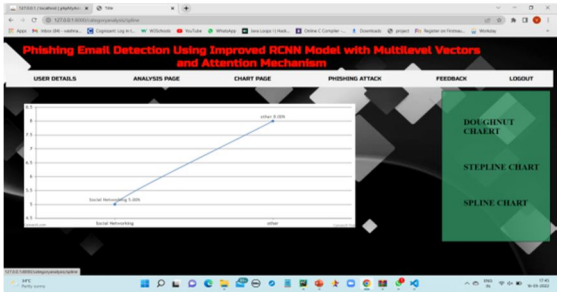


Fig 5.4 Spline Graph

VI. CONCLUSION

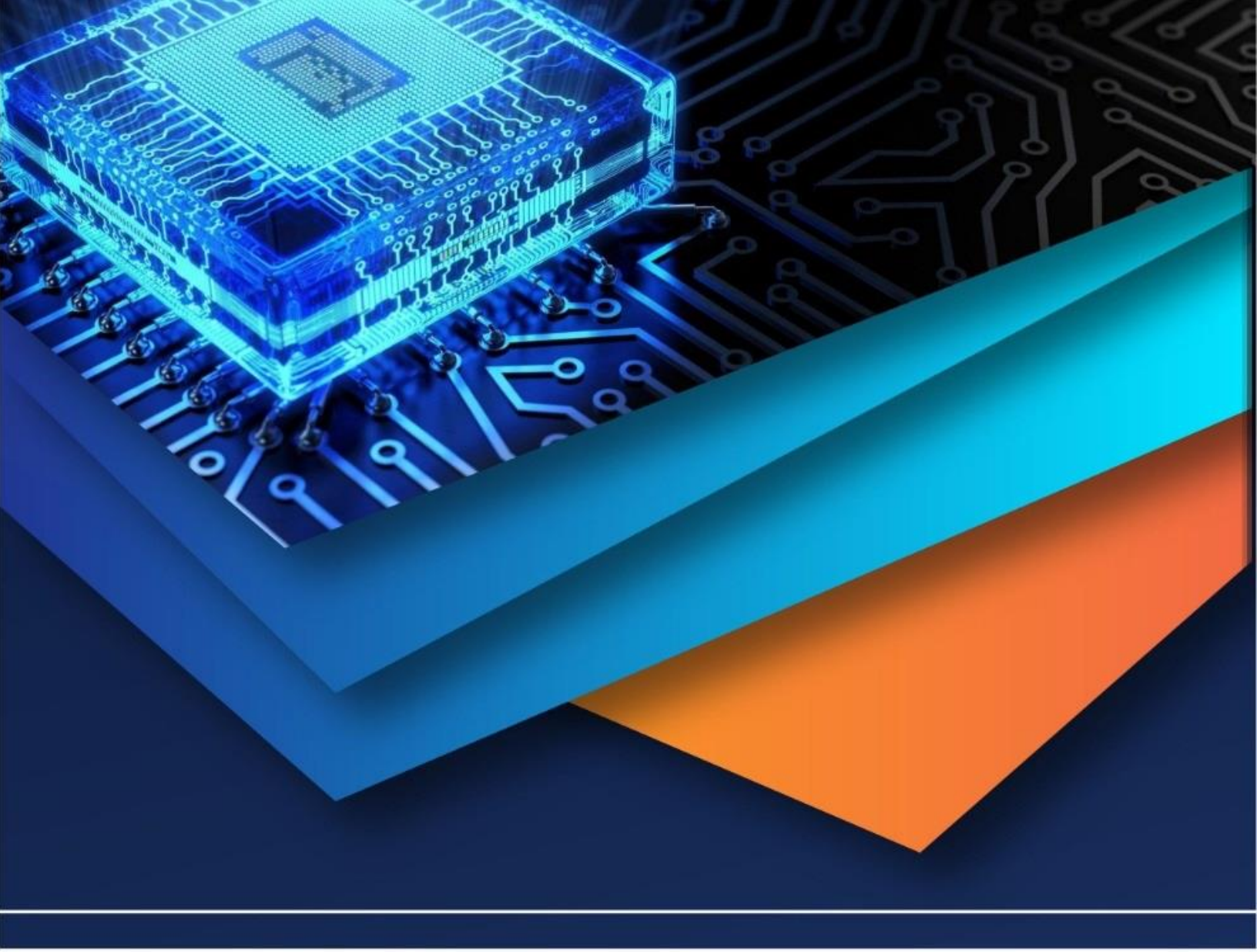
Deep Learning method is introduced to detect the advanced phishing emails . Using RCNN similar patterns from the email is recognized. Multilevel vector is used to check the email through many layers of check . Using attention mechanism different weights are assigned to different parts of the email like email header and email body to focus more on important information .To test the effectiveness of the application noise is introduced. Usage of unbalanced dataset closer to the real-world situation to conduct experiments and evaluate the model.

VII. FUTURE ENHANCEMENTS

1) An alarming trend can be adopted to detect phishing emails
2) More Phishing protections such as email security to prevent the majority of phishing attacks.

REFERENCES

[1] A.-P. W. Group et al., "Apwg attack trends report," USA: Anti-Phishing Working Group (APWG), 2014.
[2] A.-P. W. Group et al., "Phishing activity trends report 1st quarter 2018," USA: Anti-Phishing Working Group (APWG), 2018.
[3] A.-P. W. Group et al., "Phishing activity trends report 4th quarter 2016," USA: Anti-Phishing Working Group (APWG), 2017.
[4] L. M. Form, K. L. Chiew, W. K. Tiong, et al., "Phishing email detection technique by using hybrid features," in IT in Asia (CITA), 2015 9th International Conference on, pp. 1–5, IEEE, 2015.
[5] M. Nguyen, T. Nguyen, and T. H. Nguyen, "A Deep Learning Model with Hierarchical LSTMs and Supervised Attention for Anti-Phishing," arXiv preprint arXiv:1805.01554, 2018.
[6] R. Verma, N. Shashidhar, and N. Hossain, "Detecting phishing emails the natural language way," in European Symposium on Research in Computer Security, pp. 824–841, Springer, 2012.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY