International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

# Phishing Email Detection Using Improved RCNN with Multilevel

**Dr Shaik Abdul Nabi[1],Gona Srija[2], Gangula Madhuri[3], M Rohith Reddy[4,], Ch Jashuva[5]**
[2,3,4,5] UG Scholars, Department of CSE, **AVN Institute of Engineering and Technology,**Hyderabad, Telangana, India.
[1] Professor, Department of CSE, **AVN Institute of Engineering and Technology**, Hyderabad, Telangana, India.

**Abstract:**
The 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 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, 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. Experimental results show that the. 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 detecting phishing emails.

**Introduction**

Due to the increasing growth rate of internet t echnologies security has become a major conc ern for the online users . Emails are oftenly us ed to exchange data which can be personal or business related matter.The emails can contai n sensitive information which phishers want t o steal. The Phishers send a email to the recei ver which may contain link ,when receiver cli ck on that link and provide information the re ceivers information might be used for inappro priate purpose .Inorder to detect such phishing emails we are using the methods like RCNN with multilevel vectors and attention mechani sm .

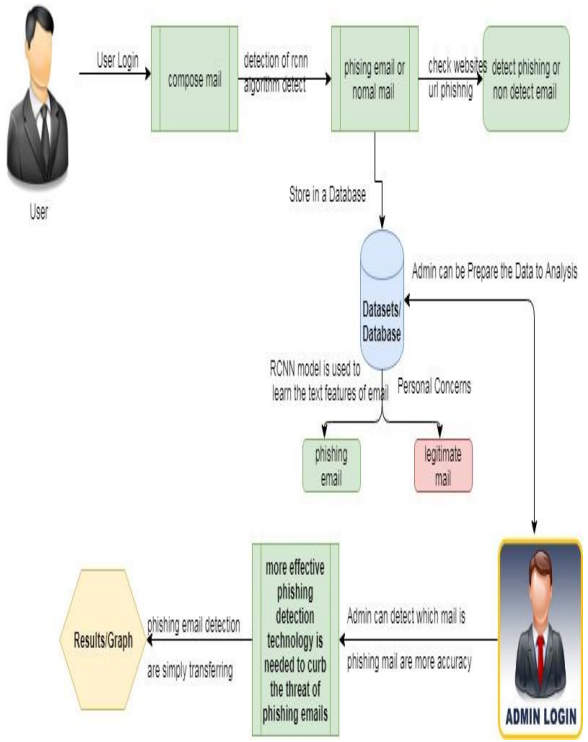
The following methods are used to detect the phishing email:

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

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

email body to focus more on important information .

**Architecture:**




**EXISTING SYSTEM:**

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 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 Given the methods mentioned above and the corresponding problems, we set to study phishing email detection systematically

based on deep learning. Specifically, this paper makes the following contributions:

**Disadvantages**

1. With respect to 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.
2. The RCNN model is improved by using the Then, the email is modelled 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.

**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

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

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.

**Advantages**

1. Phishing email refers to an attacker using a fake email to trick the recipient into returning information such as an account passwordto a designated recipient.
2. 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.

**ALGORITHM**

**R-CNN 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

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

also uses selective search for extracting the regions.

## COMPONENTS

### A. Multilevel Vector

Manual extraction of features is difficult and t ime taking ,we can not achive the effective res ults .Hence we opt for Multilevel vectors ,whi ch are very useul to extract the features from i mage or text . As in the case of Phishing emai l detection,the email has two parts namely em ail header and email body ,Multilevel vector c hecks the email header at charecter-level and word-level . It checks the email body at chare cter-level and word-level .Mostly the phishing content can be found in the email body becau se the structure of email header is mostly sam e for all the emails but the email body differs f rom email to email .The email body is more at tractive to get the attention from the victim w hich differs from legitimate mails .The words which are inappropriate and the words which t ells about the fraud or crime is detected .

### B. Attention Mechanism

Attention is an increasingly popular mechanis m used in a wide range of neural architectures .The attention mechanism is a part of a neural architecture that enables to dynamically highl ight relevant features of the input data, which, in NLP, is typically a sequence of textual ele ments. It can be applied directly to the raw inp ut or to its higher level representation. The cor e idea behind attention is to compute a weight distribution on the input sequence, assigning higher values to more relevant elements .Atte ntion can be used to compare the input data w ith a query element based on measures of simi larity or significance. It can also autonomousl y learn what is to be considered relevant, by c reating a representation encoding what the im portant data should be similar to .attention is a technique that mimics cognitive attention. Th e effect enhances some parts of the input data while diminishing other parts — the motivatio n being that the network should devote more f ocus to the small, but important, parts of the d ata. Learning which part of the data is more i mportant than another depends on the context .

### C. Neural Networks

An artificial neural network, or neural networ k, is a mathe-matical model inspired by biolog ical neural networks. Inmost cases it is an ada ptive system that changes its structure during l earning . There are many different types of N Ns. For the purpose of phishing detection, whi ch is basically a classification problem, we ch oose multilayer feedforward NN. In a feedfor ward NN, the connections between neurons d o not form a directed cycle. Contrasted with r ecurrent NNs, which are often used for pattern recognition, feedforward NNs are better at m odeling relation-ships between inputs and out puts. In our experiments, we use the most com mon structure of multilayer feedforwardNN, which consists of one input layer, one hidden layer and one output layer. The number of co mputational units in the input and output layer s corresponds to the numberof inputs and outp uts. Different numbers of units in the hidden l ayer are attempted .

### D. Deep Learning

Deep learning is a subset of machine learning, which is essentially a neural network with thr

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

ee or more layers. These neural networks atte mpt to simulate the behavior of the human bra in. ?Deep? refers to the many layers the neura l network accumulates over time, with perfor mance improving as the network gets deeper. Each level of the network processes its input d ata in a specific way, which then informs the next layer. So the output from one layer beco mes the input for the next . The adjective "dee p" in deep learning refers to the use of multipl e layers in the network. Early work showed th at a linear perceptron cannot be a universal cla ssifier, but that a network with a nonpolynomi al activation function with one hidden layer of unbounded width . Deep neural networks con sist of multiple layers of interconnected nodes , each building upon the previous layer to refi ne and optimize the prediction or categorizati on .

E. NLP

Natural Language Processing (NLP) is a subfi eld of artificial intelligence (AI). It helps mac hines process and understand the human langu age so that they can automatically perform rep etitive tasks. In natural language processing, h uman language is separated into fragments so that the grammatical structure of sentences an d the meaning of words can be analyzed and u nderstood in context. This helps computers re ad and understand spoken or written text in th e same way as humans. Email filters are one o f the most basic and initial applications of NL P online. It started out with spam filters, unco vering certain words or phrases that signal a s pam message. But filtering has upgraded, just like early adaptations of NLP. NLP combines computational linguistics—rule-based modeli ng of human language—with statistical, mach ine learning, and deep learning models. Toget her, these technologies enable computers to pr ocess human language in the form of text or v oice data and to ?understand‘ its full meaning, complete with the speaker or writer‘s intent a nd sentiment . NLP also plays a growing role i n enterprise solutions that help streamline busi ness operations, increase employee productivi ty, and simplify mission-critical business proc esses.

**MODULES:**

1. **DATASET**

The dataset has been divided into a training set and testing set. Both the training set and the testing set contain emails without header and emails with header. In this paper, we only focus on email data with the header. Due to the irrationality of the segmentation of the training set and the testing set in the original dataset, after merging the two datasets, the training-validation set and the testing set are redivided. The dataset is divided by stratified random sampling; that is, random samples are taken from legitimate email and phishing email at the same proportion. This ensures that the two datasets used in training and testing phases are well.

**2.USER QUERIES**

Users can have queries about the process. This part of project is dedicated to make and get response for queries that are needed to answerable. The major part of the modules is making project as interactive one, queries have been very normally arise to users regarding different details about the process.

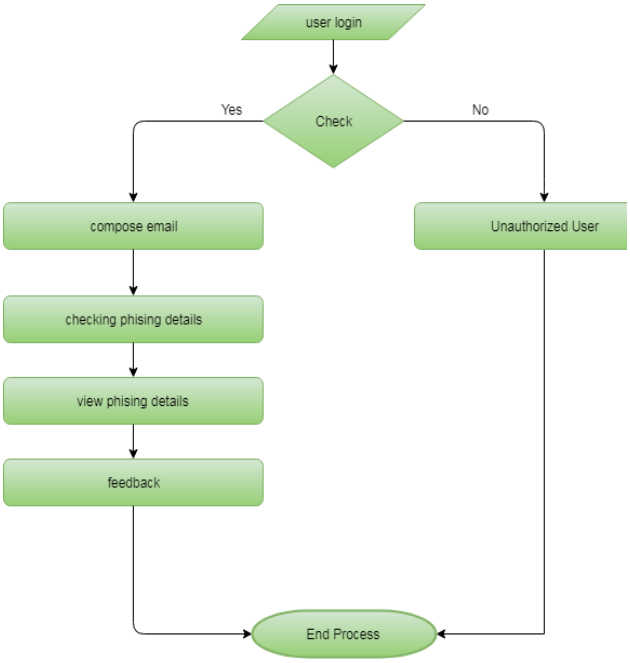
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

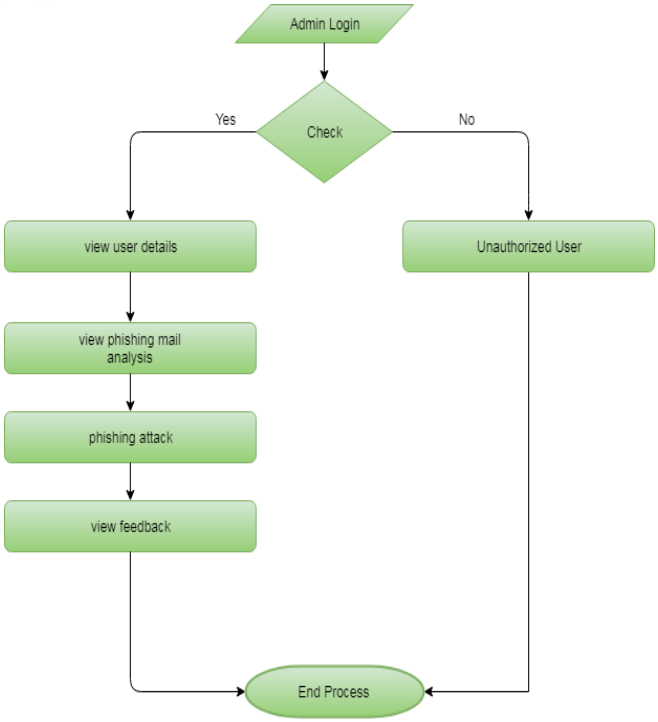
## 3.GRAPH ANALYSIS

Graph analysis is the part where admin can knows the statistics about process of details. The data are taken from the project flow and it shows until updated value. The data are gives clear solution to admin that part of improvement and user satisfaction and other factors.

## 4.ANALYSIS

analysis of email structure. a circle represents a character, and a rectangle represents a word. A rectangle is filled with an indefinite number of circles, indicating that the word consists of an indefinite number of characters.




## REQUIREMENT ANALYSIS

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

**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.

**Refernces:**

[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.