International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

Phishing Detection using Improved RCNN Model with Multilevel Vectors and Attention Mechanism

Ranjith V[1], Prathibha I Bilagi[2] ,Pritika Adhikari[3], Repshika Pradhan[4]

Assistant Professor, Dept. of CSE, RR Institute of Technology, Bengaluru, India[1]

UG Student, Dept. of CSE, RR Institute of Technology, Bengaluru, India[2,3,4]

ABSTRACT: The phishing email and Url is one of the significant threats in the today's world. Although many method of Confrontation is 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 and url.Here we first analysed the email structure. Then we proposed a new phishing email detection model, which is used to model emails at the email header, the email body, the character level, and the word level simultaneously and for Url we categories in 3 part and train in CNN-LSTM with DCDA.

KEYWORDS: Phishing, Spam Detection, Url, Email, Logistic Regression, RCNN.

I. INTRODUCTION

The phishing email and website is one of the significant threats in the world today and has caused tremendous canonical losses. Moreover, phishing emails and websites are growing at an alarming rate in recent years. Therefore, more effective phishing detection technology is needed to curb the threat of phishing emails and websites.

In this we first analyse the email and website structure. Then, based on an improved recurrent convolution neural networks (RCNN) model with multilevel vectors and attention mechanism, we proposed a new phishing email detection model named THEMIS, which is used at the email header, the email body, the character level, and the word level simultaneously. In website the character sequence features of the given URL are extracted and used for quick classification by deeplearning.

The experimental results show that the overall accuracy of THEMIS with high accuracy. Meanwhile, the false positive rate (FPR) is low. High accuracy and low FPR ensure that the alter can identify phishing emails with high probability and alter out legitimate emails as little as possible.

II. RELATEDWORK

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. Sheng et al. [10] studied the effectiveness of phishing blacklists. Blacklists mainly include sender blacklists and link blacklists. With the development of AI, phishing email detection has also entered the era of machine learning. In particular, the combination of NLP and machine learning has played a significant role in phishing email detection. Semantic features [11], syntax feature [12], and contextual features [13] previously have been used in this area. Vazhayil et al. It is obvious that both blacklisting and feature engineering need to be completed manually and require a large amount of labour and professionals with domain expertise, which limits the performance of detection. To solve the problems that exist in the first two methods, the following studies focus on deep learning techniques. Deep learning has been well-represented in many NLP tasks, including text categorization [21], information extraction [22], and machine translation [23]. It can also automatically generate effective features from emails to detect phishing emails, thus avoiding the manual extraction of email features. There are other deep learning algorithms that are being used, such as the Deep Belief Network (DBN) and the Recurrent Neural Network (RNN) [26]–[28]. At present, these deep learning methods for phishing email detection are simply transferring NLP technology to phishing email detection, ignoring the differences between phishing email detection and othertargets.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

III. PROPOSED SYSTEM

A. Design Considerations:

- Emails and website are divided into two categories, legitimate and phishing. Naturally, the detection for phishing is also a binary classification problem.
- We define a binary variable y to represent the attributes of an email or an website; that is, y = 1 means that is a phishing and y = 0 means that the email or website is legitimate.
- To determine whether the attribute is a phishing email or website, we calculate the probability that it is aspam.
- Then, the probability value is compared with the classification threshold, and if it is greater than the classification threshold, it is judged as a spam.

ADVANTAGES:

- The experimental results show that the THEMIS has high accuracy and low false positive rate(FPR)
- The approach can reduce the detection time for setting a threshold. Testing on a dataset containing millions of phishing URLs and legitimateURLs
- Our goal is to detect whether the target email or website is legitimate or phishing quickly andaccurately

B. Description of the ProposedAlgorithm:

1. MULTINOMIALNB => SPAMEMAIL

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Parameters in K neighbour classifier:

n_neighborsint, default=5 (Number of neighbors to use by default for neighbours queries).
Weights {'uniform', 'distance'} or callable, default='uniform' (Weight function used in prediction).
Possible values:

- 'uniform' : uniform weights. All points in each neighbourhood are weightedequally.
- 'distance' : weight points by the inverse of theirdistance.
- [callable] : a user-defined function which accepts an array of distances, and returns an array of the same shape containing the weights.

Algorithm {'auto', 'ball_tree', 'kd_tree', 'brute'}, default='auto'

Attributes:

- **classes_array of shape (n_classes,):** Class labels known to the classifier
- **effective_metric_str or callable:** The distance metric used. It will be same as the metric parameter or a synonym of it, e.g. 'euclidean' if the metric parameter set to 'minkowski' and p parameter set to2.
- **effective_metric_params_dict:** Additional keyword arguments for the metric function. For most metrics will be same with metric_params parameter, but may also contain the p parameter valueif the effective_metric_ attribute is set to 'minkowski'.
- **outputs_2d_bool:** False when y's shape is (n_samples,) or (n_samples, 1) during fit otherwiseTrue.

2. Logistic Regression =>URL Spam

This class implements regularized logistic regression using the 'liblinear' library, 'newton-cg', 'sag', 'saga' and 'lbfgs' solvers.

The 'liblinear' solver supports both L1 and L2 regularization, with a dual formulation only for the L2 penalty. The Elastic-Net regularization is only supported by the 'saga' solver.

Attributes:

- **classes_ndarray of shape (n_classes,) :** A list of class labels known to theclassifier.
- **coef_ndarray of shape (1, n_features) or (n_classes, n_features) :** Coefficient of the features in the decision function.
- **coef_ is of shape (1, n_features) :** when the given problem is binary. Inparticular, when multi_class='multinomial', coef_ corresponds to outcome 1 (True) and -coef_ corresponds to outcome 0

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

(False).

- **intercept_ndarray of shape (1,) or (n_classes,) :** Intercept (bias) added to the decisionfunction.
- If **fit_intercept intercept** is set to False, the intercept is set to zero. intercept_ is of shape when the given problem is binary. In particular, when muliti_class='multinomial', intercept_ corresponds to outcome 1 (True) and -intercept_ corresponds to outcome 0(False).
- **n_iter_ndarray of shape (n_classes,) or (1,) :** Actual number of iterations for all classes. If binary or multinomial, it returns only 1 element. For liblinear solver, only the maximum number of iteration acrossall classes isgiven.

IV. PSEUDOCODE

Step 1: Access Data set from pandas and convert it to frame
Step 2: Convert the data to Vector numpy array form
Step 3: Split the data to training and testing set
Step 4: Train the model by Fit method
Step 5:

1. LogisticRegression
   lgr=LogisticRegression() //SPAM URL Prediction
   lgr.fit(x_train,y_train)

Step 6:

2. MultinomialNB //SPAM EmailPrediction
   clf = MultinomialNB()
   clf.fit(X_train,y_train)

Step 7: Predict the result

System Design

a) Data Flow Diagram

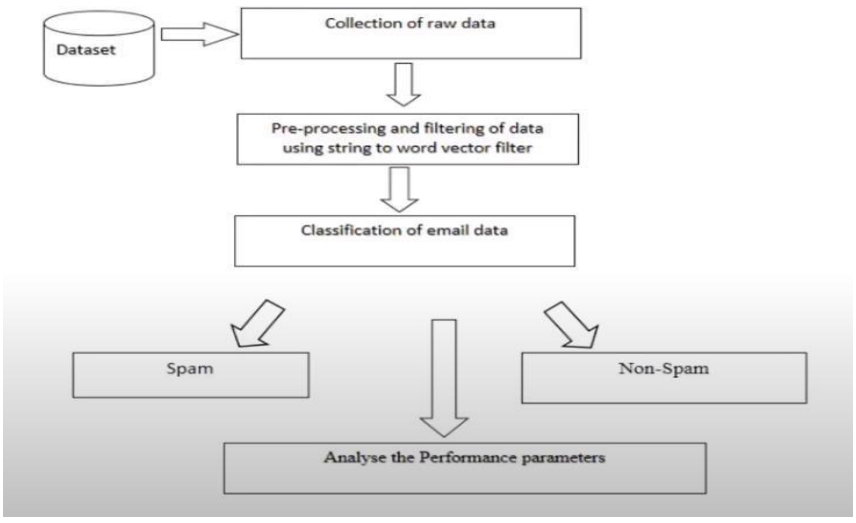

Fig 1: Data Flow Diagram

This shows the data flow diagram of phishing detection model. And the test case is shown below:

- Spam Email

  Test case 1: Test the model with real world spam email
  Actual output: Flags the email as a spam

  Expected output: Email is spam

  Example Email: WINNER!! As a valued network customer you have been selected to received å£900

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only

Test case 2: Test the model with genuine Email
Actual output: Flags the email as Not a Spam
Expected output: Email is not spam
Example Email : o until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat

- Spam Url

  Test case 1: Test the model with real world spam Url
  Actual output: Flags the Url as a spam
  Expected output: Url is spam
  Example Url: diaryofagameaddict.com

  Test case 2: Test the model with genuine Url
  Actual output: Flags the Url as Not a Spam
  Expected output: Url is not spam
  Example Url: facebook.com

V. SIMULATIONRESULTS

In this we use a new deep learning model to detect phishing emails and phishing website detection solutions with sentimental analysis and logistic regression. The model employs an improved RCNN to model the email header and the email body at both the character level and the word level with analyzing the content and url of the website.

We conduct a series of experiments on a dataset containing millions of phishing and legitimate emails and URLs. From the results, we find that the approach of this model is effective with high accuracy, low false positive rate and high detection speed.

1. Home Page

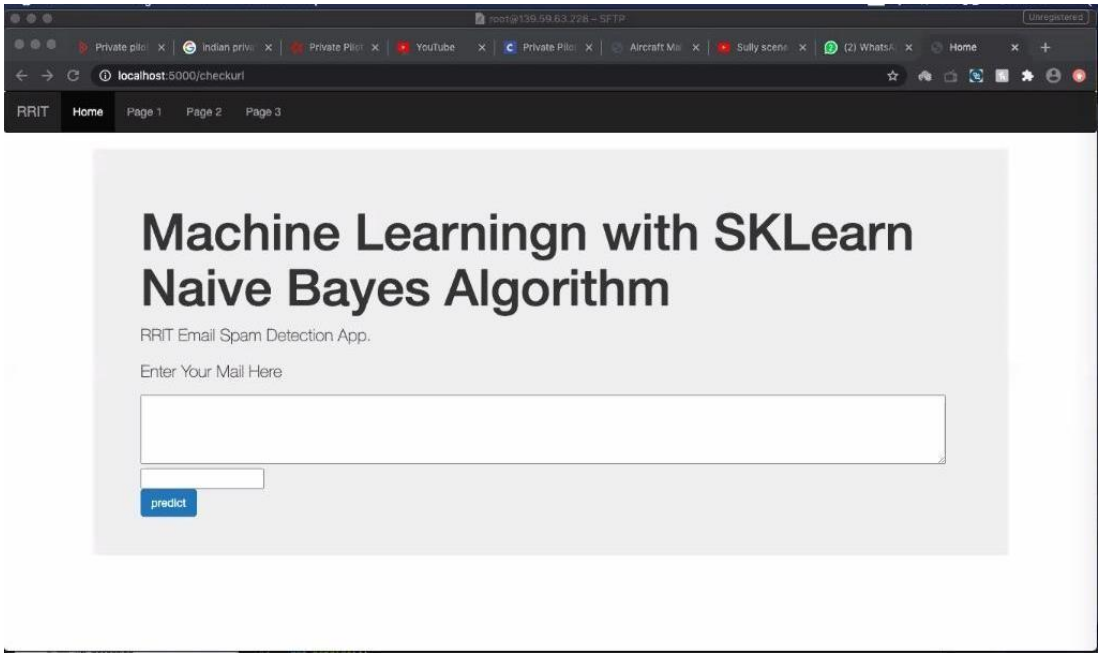

Fig 2: Home Page

2. Text data or file upload has input

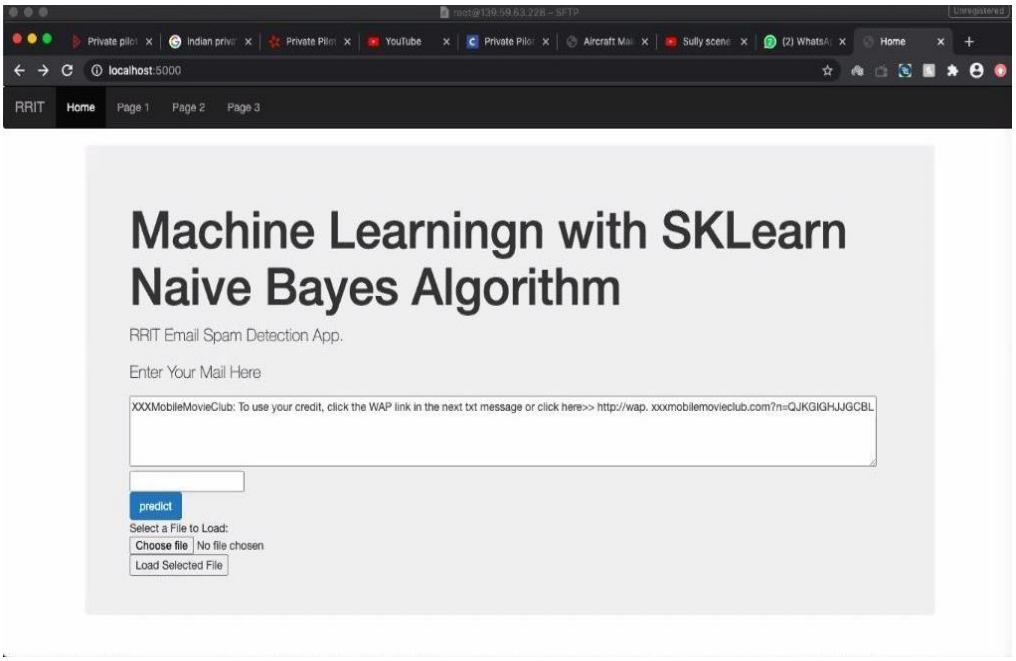

Fig 3: Email or text file has input

3. URL hasinput

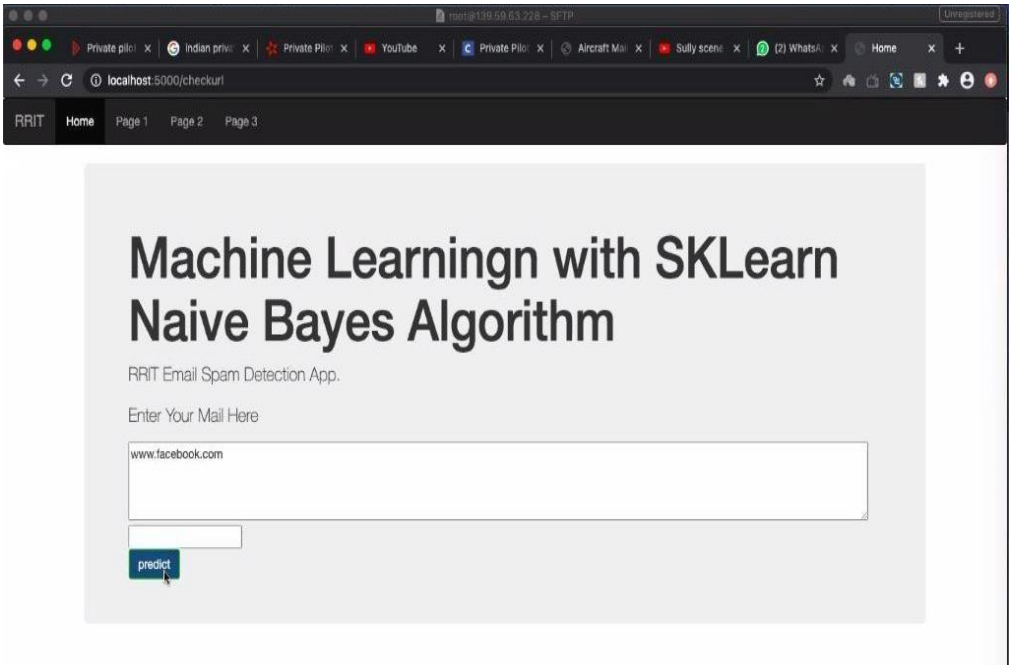

Fig 4: URL has Input

4. Final Result

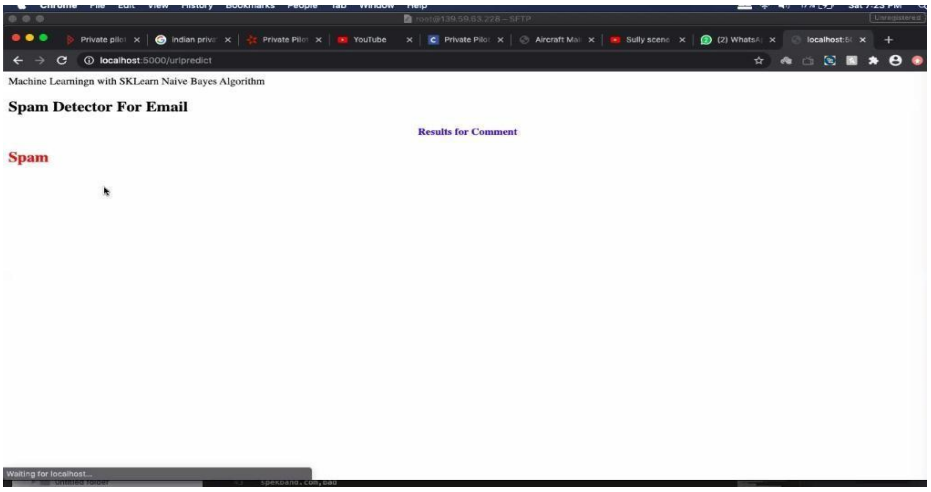

Fig 5: Result Page

VI. CONCLUSION AND FUTURE WORK

In this paper, 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. Then, Word2Vec 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. The CNN-LSTM algorithm contains three parts as URL embedded representation, feature extraction, classification. Therefore, the noise is introduced into the model minimally. In the model, we use the attention mechanism and logistic regression. Here 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.

REFERENCES

1. N. Lord, “What is a Phishing Attack? Defining and Identifying Different Types of Phishing Attacks”. https://digitalguardian.com/blog/whatphishing-attack-definingand-identifying-different-types-phishingattacks, 2018..
2. N. Sadeh, A. Tomasic, and I Fette, “Learning to detect phishing emails”, Proceedings ofthe16thinternational conference on world wide web, pp.649–656,2007.
3. J. Ma, S. S. Savag, G. M. Voelker, “Learning to detect malicious URLs”, ACM Transactions on Intelligent Systems and technology, vol. 2, no. 9, pp 30:1-30:24,2011.
4. S. Purkait, “Phishing counter measures and their effectiveness–literature review”, Information Management & Computer Security, vol. 20, no. 5, pp. 382–420,2012.
5. N. Abdelhamid, A. Ayesh, F. Thabtah, “Phishing Detection based Associative Classification”, Data Mining. Expert Systems with Applications (ESWA), vol. 41, pp 5948-5959,2014.
6. D.R Patil and J. Patil, J., “Survey on malicious web pages detection techniques”, International Journal of u- and e-Service, Science and Technology, vol. 8, no. 5, pp. 195–206,2015.
7. W. Hadi, F. Aburrub, and S, Alhawari, “A new fast associative classification algorithm for detecting phishing websites”, Applied Soft Computing vol. 48, pp 729- 734,2016.
8. R. K. Nepali and Y. Wang, Y., “You look suspicious!! Leveraging visible attributes to classify malicious short urls on twitter”, 2016 49th Hawaii International Conference on System Sciences (HICSS). IEEE, pp. 2648–2655,2016.

BIOGRAPHY

Prof. Ranjith V is an Assistant Professor in the Computer Science & Engineering Department, R.R Institute of Technology, Visvesvaraya Technological University. He received BE and Master of Technology degree and PhD (pursuing). His area of interest is cloud and cloud security.