iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6093

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: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Classification of E-mail (Phishy or Ham)

Jaydip Nakarani[1], Ajay Vandra[2], Aayush Vaishnav[3], Ayush Trivedi[4], Atul Kumar[5]
[1, 2, 3, 4]*B.Tech. Student,* [5]*Assistant Professor, Dept. of Computer Science and Engineering, Parul University, Waghodia Road, Limda, Vadodara, INDIA*

Abstract: ***Nowadays cyber-attacks are increase as compare to last 5 years. Hacker can attack on device through message, tricky image, e-mail etc. last few years hackers have done many cyber-attacks through text message and image. They send email to recipient (user) and tricked to click on malicious link, which can lead to the installation of malware, the freezing of the system as part of a ransom ware attack or the revealing of sensitive information. We want to classify this type of mail into whether it is Phishy or Ham. Our project aim is to classify received E-mail is Phishy E-mail or Ham. Machine learning algorithms and techniques used for predict received E-mail is Phishy or Ham. We are design one model which take input as raw data (M-BOX file which is contain text file of more than one E-mail) and predefined label and gives prediction of E-mail. We will use Supervised learning algorithm and Classification algorithm like Naïve Bayes, Support vector machine, Random forest, Decision tree etc.***

I. INTRODUCTION

Email (Electronic Mail) is one of the efficient ways to exchange data in the current century. it is very effective to share data all over the world. People are sharing personal information or documents through email, so it is very important for user to their data is safe. Cyber-attacks are increases for last few years. Cyber-attackers send unwanted commercial bulk emails and create a huge problem on internet. Cyber-attackers gather email addresses from different sources like websites, social platform etc. cyber-attackers send email which has malicious links, viruses etc. which can freeze system as part of ransomware attack and revealing of sensitive information. this type of email steals your personal information, like password bank verification number, credit card number etc. This phishy email creates many problems. phishy email interrupt the business productivity. Phishy email has pattern to confuse the end user and steal their information. Phishy email carry malicious links in a form of picture or a zip file or document. if the user downloads the picture or a file. malware was activated in your mobile or a computer and steal user's information. Phishy email consume the lot of bandwidth space. if we can stop this phishy emails we can save lot of bandwidth. phishy email have warning sign like urgent offers (for example “Buy now and get 50% off”). so, it's necessary to stop this kind of emails. We have to classify these emails, so we use machine learning techniques. We classify the email is phishy or ham. Phishy email means which has malicious links. ham email is a normal(safe) email. Using machine learning first we have to train model for classification. For classification we use naive bayes (NB), random forest (RF), decision tree, support vector machine (SVM). We use various machine learning algorithm for find phishy email with high accuracy model. We have start from pre-processing and cleaning email format. Such as removing unnecessary words, stemming. Then we compressing feature and finally implement feature selection techniques. Our main target is to preserve most important feature not all features.

II. LITERATURE REVIEW

Information is interchange by heavily use of E-mail. There are many chances to get fraud e-mail by any of user and which is harmful for our security and it can lead to steal our personal information from our device like card details, credentials, bank information, personal information etc, this kind of mail called as Phishy [2] E-mail. Opposite from this a mail which is not harmful for our information and security it’s called Ham [2] E-mail. So, we have to classify this mail and inform end user to phishy e-mail is dangerous for your system. One survey has been done on discuss methods of protection against phishing email attacks in detail [4]. They present an overview to use Machine Learning (ML) [6] technique to protect against phishing email. Most classifiers used to identify phishing email are based on supervised learning. Supervised learning technique build one model which is first train and after used to detect phishing e-mail based on given data set of e-mail with predefined label. There is various algorithm for classification available in literature varied from accuracy, performance and size of dataset. Classification on e-mail has been done through spam and non-spam e-mail. Paper on (Machine learning for email spam filtering: review, approaches and open research problems), used Machine Learning (ML) technique for classification. There are many techniques for classification like supervised and unsupervised learning in machine learning. For supervised learning many algorithm available like Naïve Bayes (NB) [1], Decision Tree [1], Random Forest [1], Support Vector Machine (SVM) [1] etc, based on which dataset they are used.

Classification can be more accurate if we are used best algorithm but result may vary upon size of dataset. In case of classification most of problem classify based on text patent [7]. First of all, we have to identify text patent from e-mail body and remove stop word from text. For classification of mail into spam [1] and non-spam [1] 27 features is used. E.g. 1. Message format, 2. Number of URLs, 3. Keywords. These are the features used for classification of e-mail into phishy [2] or ham [2].

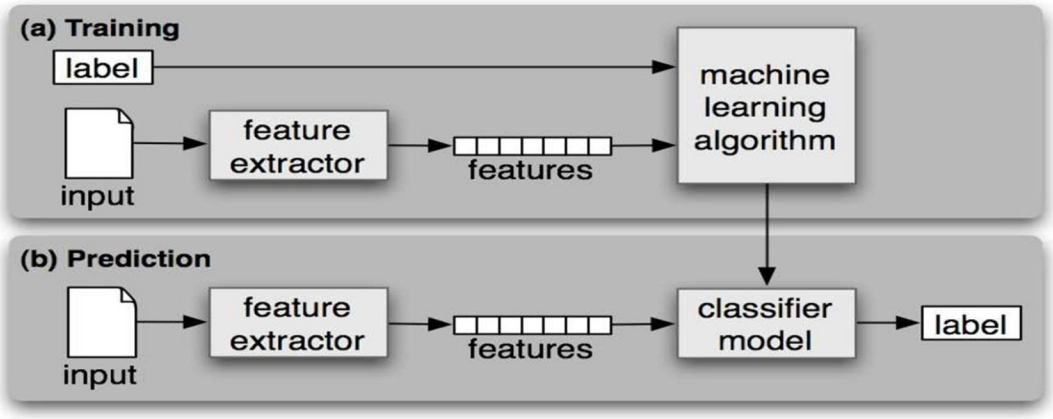


Figure 1. Work flow of classification [1][6]

In supervised learning technique there is given predefined label and input form text file (M-box a file which is contain more than one e-mail text) from there need to extract features which is select by feature selection [4]. After process of feature selection machine learning algorithm apply on them and build one classifier model which is able to predict whether it is phishy [3] or ham [3] e-mail.

III. PROPOSED METHODOLOGY

Most of filtering method used text techniques so all problems linked with classification. Our research present extract features from e-mails and removed unnecessary features. Various Machine Learning classification algorithm are presents today but we are focus on only four algorithms which is Naïve Bayes (NB), Decision Tree, Random Forest, Support Vector Machine (SVM). We use only supervise learning method because we have used discrete data set for classification.

A. Naïve Bayes (NB)

Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Naïve Bayes classifier works on following formula:

Likelihood — Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability — Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 2. Formula of bayes algorithm

Where,

1) P(c|x) is the posterior probability of class (target) given predictor (attributes).
2) P(c) is the prior probability of class.
3) P(x|c) is the likelihood which is the probability of predictor given class.
4) P(x) is the prior probability of predictor.

Naïve bayes used for find probability of data which is available by decision or not. Posterior probability finds by particular instance available with product to total instance available in same class divide by total number of availabilities. Naïve bayes is very useful for classification where classification is based on probability.

B. Decision Tree

Decision tree classifier used to classify data into respective class which they are belongs to class. Decision tree worked on splitting data into sub-nodes start with root node. Where leaf node called as which not splitting further. We have extract features from the e-mail which is splitting into sub-node if the feature set meets to phishy feature set then classify into phishy else ham. Decision trees classify the examples by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the example. Each node in the tree acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. This process is recursive in nature and is repeated for every subtree rooted at the new nodes. Provide a clear indication of which fields are most important for prediction or classification.

C. Random Forest

Random forest is an ensemble classifier which is made using many decision tree classifiers. Model which is combine results from different models. A huge number of unrelated trees functions like one model and outcome is individual with respect to classification of every tree. Random forest works like decision tree but difference is in random forest use many trees to classify into respective class. Feature set already provide with classifier so random forest classifier made a decision respect to which feature set is belongs to phishy class or ham class. We have used 40 features for classify e-mail into whether it is phishy or ham.

D. Support Vector Machine (SVM)

Support vector machine is based on the idea of finding a hyperplane that best divides a dataset into two classes with support vector. Support vector is participant data which is classify into separate class further. SVM works on find nearest hyperplane which is divided support vector into their respective class. Hyperplane limited as boundary of support vectors. Hyperplane basically work on 2D view where data is clearly far from hyperplane but what if data is too dense? So, in case we have to use 3D view to classify data into respective class. SVM is more accurate and efficient on small data size.

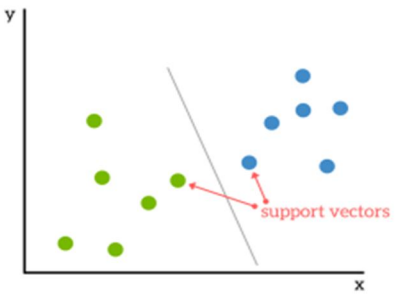


Figure 3. support vectors

E. Feature Extraction

Feature extraction algorithm extracts features from e-mails data set. We have used 40 features and apply algorithm to classify e-mail. List of features is following:

1) Body_formsS
2) Body_html
3) Body_noCharacters
4) Body_noDistinctWords
5) Body_noFunctionWords
6) Body_noWords
7) Body_richness
8) Body_suspension

These are the features which is extracted from E-mails which is used by machine learning (ML) algorithm for classify e-mail whether it is phishy or ham.

IV. RESULTS AND OBSERVATIONS

We have tested four algorithm and choose high accurate algorithm for classify e-mail into phishy or ham. Result may vary upon size of dataset. We have tested 3600 e-mail for testing purpose and find accuracy of every algorithm. We observe that random forest has highest accuracy as 99.05% followed by decision tree which has 98.77% accuracy. This accuracy may vary upon quantity of dataset. Fig 4,5,6 and 7 shows accuracy of model respective Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM) and Random Forest (RF).

```
[[1072   27]
 [  82  933]]
              precision    recall  f1-score   support

           0       0.93      0.98      0.95      1099
           1       0.97      0.92      0.94      1015

    accuracy                           0.95      2114
   macro avg       0.95      0.95      0.95      2114
weighted avg       0.95      0.95      0.95      2114

Accuracy is: 94.84389782403028
============================
Fold-1 Accurecy- 0.9801699716713881
Fold-2 Accurecy- 0.924822695035461
Fold-3 Accurecy- 0.924822695035461
Fold-4 Accurecy- 0.9318181818181818
Fold-5 Accurecy- 0.9375
Fold-6 Accurecy- 0.9744318181818182
Fold-7 Accurecy- 0.9744318181818182
Fold-8 Accurecy- 0.9545454545454546
Fold-9 Accurecy- 0.8877840909090909
Fold-10 Accurecy- 0.9659090909090909
mean accurecy-0.9456235816287766
============================
Fold-1 F1-score- 0.9801621708413616
Fold-2 F1-score- 0.9243790337759483
Fold-3 F1-score- 0.9247039241013942
Fold-4 F1-score- 0.9318044220570072
Fold-5 F1-score- 0.9373706004140787
Fold-6 F1-score- 0.9744217027976263
Fold-7 F1-score- 0.9743968965935265
Fold-8 F1-score- 0.9543982381137452
Fold-9 F1-score- 0.88620231744921
Fold-10 F1-score- 0.9658550780871085
mean f1-score-0.9453694384231005
```

Figure 4. output of NB

```
[[700   5]
 [ 13 691]]
              precision    recall  f1-score   support

           H       0.98      0.99      0.99       705
           P       0.99      0.98      0.99       704

    accuracy                           0.99      1409
   macro avg       0.99      0.99      0.99      1409
weighted avg       0.99      0.99      0.99      1409

Accuracy is: 98.72249822569198
============================
Fold-1 Accurecy- 0.9929178470254958
Fold-2 Accurecy- 0.9645390070921985
Fold-3 Accurecy- 0.9858156028368794
Fold-4 Accurecy- 0.96875
Fold-5 Accurecy- 0.9701704545454546
Fold-6 Accurecy- 0.9957386363636364
Fold-7 Accurecy- 0.9928977272727273
Fold-8 Accurecy- 0.9815340909090909
Fold-9 Accurecy- 0.9900568181818182
Fold-10 Accurecy- 0.9928977272727273
mean accurecy-0.9835317911500028
============================
Fold-1 F1-score- 0.9929177191444284
Fold-2 F1-score- 0.9702106082830807
Fold-3 F1-score- 0.988651363446122
Fold-4 F1-score- 0.9687459640966034
Fold-5 F1-score- 0.9701385817038217
Fold-6 F1-score- 0.9971588845393278
Fold-7 F1-score- 0.99431653050021
Fold-8 F1-score- 0.9872084501753388
Fold-9 F1-score- 0.9914739009325422
Fold-10 F1-score- 0.992896566337393
mean f1-score-0.9853718569158868
```

Figure 5. output of DT

```
[[1029   15]
 [  42 1028]]
              precision    recall  f1-score   support

           H       0.96      0.99      0.97      1044
           P       0.99      0.96      0.97      1070

    accuracy                           0.97      2114
   macro avg       0.97      0.97      0.97      2114
weighted avg       0.97      0.97      0.97      2114

Accuracy is: 97.30368968779565
============================
Fold-1 Accurecy- 0.9759206798866855
Fold-2 Accurecy- 0.9418439716312057
Fold-3 Accurecy- 0.9702127659574468
Fold-4 Accurecy- 0.9644886363636364
Fold-5 Accurecy- 0.9389204545454546
Fold-6 Accurecy- 0.9801136363636364
Fold-7 Accurecy- 0.9900568181818182
Fold-8 Accurecy- 0.9829545454545454
Fold-9 Accurecy- 0.9829545454545454
Fold-10 Accurecy- 0.9857954545454546
mean accurecy-0.9713261508384429
============================
Fold-1 F1-score- 0.9759167661606687
Fold-2 F1-score- 0.9417083947574862
Fold-3 F1-score- 0.9702041333840496
Fold-4 F1-score- 0.9644851251193078
Fold-5 F1-score- 0.938884817504396
Fold-6 F1-score- 0.9801078567507346
Fold-7 F1-score- 0.9900534265017247
Fold-8 F1-score- 0.982940777357759
Fold-9 F1-score- 0.9829511055142518
Fold-10 F1-score- 0.9857936200439106
mean f1-score-0.9713046023094289
```

Figure 6 output of SVM

```
[[1074    4]
 [  16 1020]]
              precision    recall  f1-score   support

           H       0.99      1.00      0.99      1078
           P       1.00      0.98      0.99      1036

    accuracy                           0.99      2114
   macro avg       0.99      0.99      0.99      2114
weighted avg       0.99      0.99      0.99      2114

Accuracy is: 99.05392620624409
============================
Fold-1 Accurecy- 0.9915014164305949
Fold-2 Accurecy- 0.9801418439716312
Fold-3 Accurecy- 0.9858156028368794
Fold-4 Accurecy- 0.9786931818181818
Fold-5 Accurecy- 0.9758522727272727
Fold-6 Accurecy- 0.9985795454545454
Fold-7 Accurecy- 0.9957386363636364
Fold-8 Accurecy- 0.9886363636363636
Fold-9 Accurecy- 0.9914772727272727
Fold-10 Accurecy- 0.9957386363636364
mean accurecy-0.9882174772330016
============================
Fold-1 F1-score- 0.991501348228043
Fold-2 F1-score- 0.9801302905366231
Fold-3 F1-score- 0.9858073503833021
Fold-4 F1-score- 0.9786927948959558
Fold-5 F1-score- 0.9758307669005907
Fold-6 F1-score- 0.9985794050047723
Fold-7 F1-score- 0.9957375957375958
Fold-8 F1-score- 0.9886304909560724
Fold-9 F1-score- 0.9914747957503148
Fold-10 F1-score- 0.9957379398024357
mean f1-score-0.9882122778195706
```

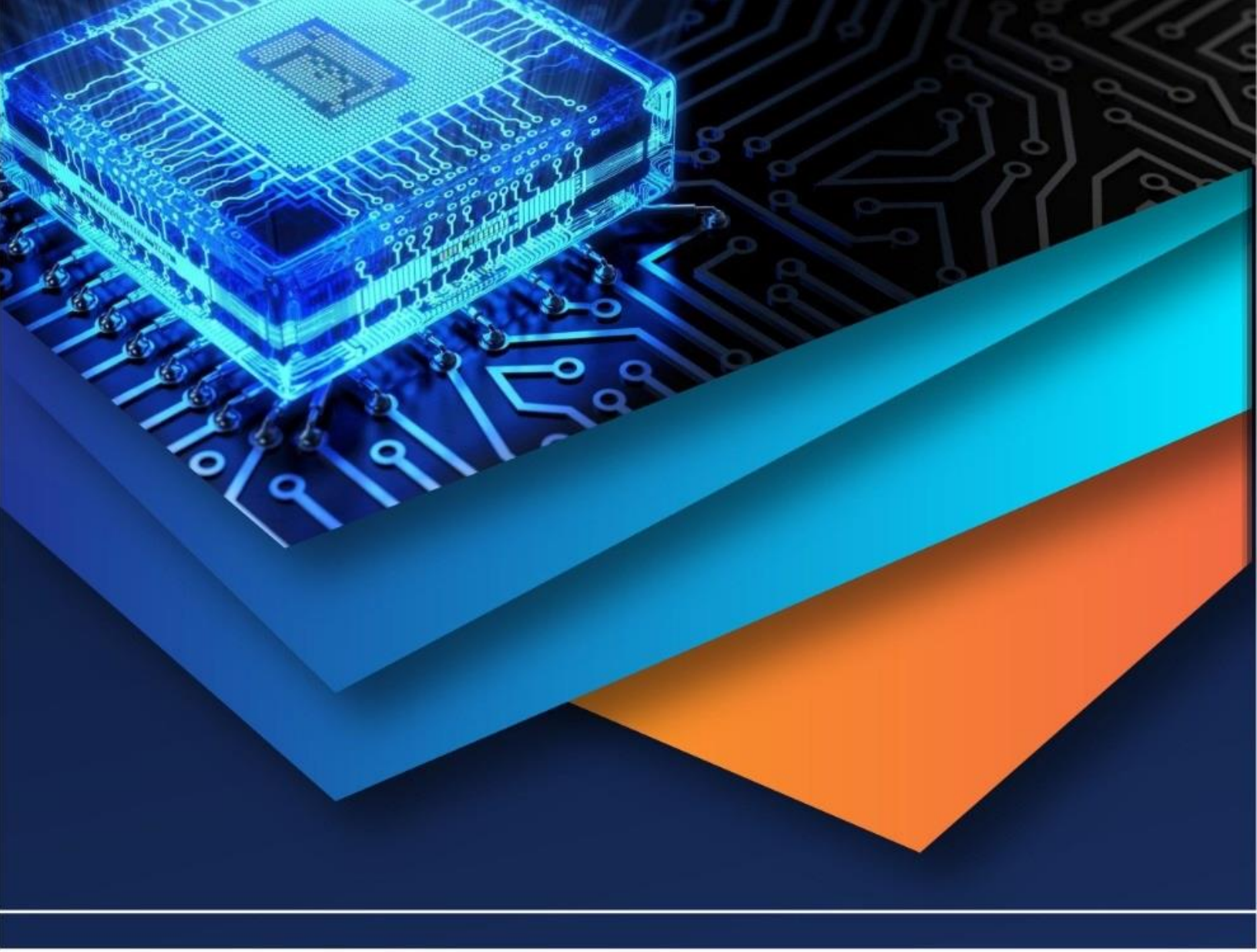
Figure 7 output of RF

V. CONCLUSION AND FUTURE WORK

In the future, we will try to test more e-mail so we will get perfect accuracy of model. The technique proposed here give results based on accuracy and F1 score. F1 score measured with True Positive, True Negative, False Positive and False Negative. We have chosen Random Forest algorithm (with Accuracy 99.05%) for classification whether it is phishy or ham. In future we can integrate our proposed method to Google, Yahoo, Microsoft etc, mail service provider with more specific accuracy.

REFRENCES

[1] Uysal Halper Kursat, “An improved global feature selection scheme for text classification”, Expert systems with applications 43(2016), pp. 82-92
[2] Abdelhamid Neda, Thabtah Fadi, Abdel-Jaber Hussein, “Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features”, IEEE (2017), pp. 71-77
[3] Fernando Sanchez, Zhenhai Duan, “A Sender-Centric Approach to Detecting Phishing Emails”, IEEE (2012), pp.32-39
[4] Rajput Amandeep, Sohal J S, Athavale Vijay, “Email header feature extraction using adaptive and collaborative approach for Email classification, IJITEE (2019), Volume-8, Issue-7S, pp.158-164
[5] Ammar Almomani, B. B. Gupta, Samer Atawneh, A. Meulenberg, Eman Almomani, “A Survey of Phishing Email Filtering Techniques”, IEEE (2013), Volume-15, pp.2070-2090
[6] Naghmeh Moradpoor, Benjamin Clavie, Bill Buchanan, “Employing Machine Learning Techniques for Detection and Classification of Phishing Emails”, IEEE (2017), pp.149-156
[7] Fujin Zhu, Xuefeng wang, Donghua Zhu, Yuquin Liu, “A Supervised Requirement-oriented Classification Scheme Based on Combination of Meta data And Citation Information”, International Journal of Computationalntelligence Systems (2015), pp.502-516



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY