iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35226

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 9 Issue VI Jun 2021- Available at www.ijraset.com

Prediction of Spam Email using Machine Learning Classification Algorithm

P Sai Teja[1], Ch. Amith[2], K. Deepika[3], Dr. K. Srujan Raju[4]
[1, 2, 3]*B.Tech 4th year, Department of Computer Science and Engineering, CMR Technical Campus*
[4]*Professor & Head of Department, Department of Computer Science and Engineering, CMR Technical Campus*

Abstract: ***Unsolicited e-mail also known as Spam has become a huge concern for each e-mail user. In recent times, it is very difficult to filter spam emails as these emails are produced or created or written in a very special manner so that anti-spam filters cannot detect such emails. This paper compares and reviews performance metrics of certain categories of supervised machine learning techniques such as SVM (Support Vector Machine), Random Forest, Decision Tree, CNN, (Convolutional Neural Network), KNN(K Nearest Neighbor), MLP(Multi-Layer Perceptron), Adaboost (Adaptive Boosting) Naïve Bayes algorithm to predict or classify into spam emails. The objective of this study is to consider the details or content of the emails, learn a finite dataset available and to develop a classification model that will be able to predict or classify whether an e-mail is spam or not.***

Keyword: Spam Email, Classification, Dataset, Performance Metrics.

I. INTRODUCTION

Spam email is unsolicited and unwanted junk email sent out in massive amount or in bulk to an indiscriminate recipient list. Generally, spam is sent for commercial purposes[6]. It is sent in massive volume by botnets, networks of infected computers. Spam email can often be a malicious attempt to gain access to your system. Spam prevents the user from making full and good utilization of cpu time, storage capacity and network bandwidth. It becomes a huge problem especially at times when there are Spam mails which come in between important business mails. Hence, it becomes inevitable to solve such problems which are encountered by spam email. So, this problem can be solved by using Machine Learning methods which can successfully detect and filter spam. It is also important to find out which technique or algorithm can best fit in the purpose of classifying spam mail.

II. PROBLEM STATEMENT

The person responsible for sending the spam messages is referred to as the spammer. Such a person gathers email addresses from different websites, chat rooms etc. The huge volume of Spam mails flowing through the computer networks have destructive effects on the memory space of the email server, communication bandwith, cpu power and user time. In all, existing system does not find spam mails effectively. Hence, it also results in untold financial losses to many users. It leads to low test and prediction accuracy, less security and also loss of data.

III. LITERATURE SURVEY

The field of Machine Learning includes several algorithms which result in their own particular levels of accuracy. At present, several experts are working in the email classification domain to classify an email into ham or spam or into phishing or legitimate. However, very few review studies are available in the literature on spam email classification and phishing email classification from the text classification perspective.

For example, Blanzieri and Bryl [4] presented a structured overview of existing learning-based approaches to spam filtering. In addition, a survey on datasets, text- and image-based features, performance measures, and spam filtering algorithms was presented. Guzella and Caminhas [5] investigated the available datasets, feature reduction techniques, and classification algorithms to identify spam emails.

They also examined the literature on image-based spam email classification. Although both reviews are on spam email classification, they are outdated. In earlier research by AK Sharma, S Sahni [2]. Machine Learning algorithms such as ID3,J48, SimpleCart, ADTree were used. These algorithms produced the following accuracy results.[2]

Instances(4601) / Algorithms	Correctly classified instances	Incorrectly classified instances
ID3	4100(89.1111%)	433(9.411%)
J48	4268(92.7624%)	333 (7.2376%)
ADTree	4183(90.915%)	418 (9.085%)
SimpleCART	4262(92.632%)	339(7.368%)

Figure 1-Accuracy metrics

Here, we observe that the above algorithms produced accuracy of upto maximum 92%.In the proposed methodology ,below we are using AdaBoost and other algorithms which provide accuracy of about 93-94%.

IV. RELATED WORK

Email is one of the most popular and frequently used ways of communicating due to its worldwide accessibility, relatively fast message transfer, and low sending cost .The flaw in email protocols and increasing amount of electronic business and financial transactions directly contribute to increasing in email based threats. The growth of email users has resulted in the dramatic increasing of spam emails during past few years. Spam in emails has become major issue. Spam messages consume space, network bandwidth & are of no use to receiver. It is very difficult to filter spam as spammers try to tackle processes carried out by filtering mechanism. This serious issue has generated need for efficient and effective anti spam filters that filter email into spam or ham email. Spam filters prevent spam emails from getting into users inbox. Email spam filters can filter emails on the basis of content base or on header base. Various spam filters are labeled into 2 categories Machine Learning and Non-Machine Learning. Machine learning algorithm in the area of spam filtering is most commonly used. Machine Learning plays a crucial role to render spam filtering more efficiently.

V. PROPOSED METHODOLOGY

In this paper, we are giving brief review on various machine learning algorithms such as SVM (Support Vector Machine), Random Forest, Decision Tree, CNN (Convolutional Neural Network), KNN(K Nearest Neighbor), MLP(Multi-Layer Perceptron), Adaboost (Adaptive Boosting) ,Naïve Bayes algorithm to predict or classify into spam emails**.** There are various SPAM datasets such as SPAM ARCHIVE, SPAMBASE, LINGSPAM etc to perform this experiment. We are using SPAMBASE dataset to evaluate performance of above algorithms in terms of Accuracy,Precision and Recall.

1) *SVM:* It refers to support vector machine. It is basically a hyperplane which classifies data into classes.[9]
2) *Decision Tree:* It uses tree representation to solve the problem in which leaf node corresponds to a class label and attributes are represented as internal nodes.[7]
3) *AdaBoost:* It is short for Adaptive Boosting and is a very popular boosting technique which combines multiple weak classifiers into single strong classifier.
4) *KNN:* It refers to k-nearest neighbor. It is a simple algorithm which performs classification based on similarity measure.
5) *Naïve Bayes:* It is a supervised learning algorithm, which is based on Bayes theorem which is a probablistic classifier[11] .
6) *MLP:* It refers to multi layer perceptron. It is formed from multiple layers of perceptron. It consists of three layers input,hidden and output layers.[8]
7) *Random Forest:* It is a classifier that contains a number of decision trees on various subsets of given dataset and takes the average to improve the predictive accuracy of that dataset.
8) *CNN:* It refers to convolutional neural network.It is similar to MLP but is more effective .It has much deeper layers and are sparsely connected rather than fully connected.

A. *Advantages*

1) Security is more.
2) Accuracy is more.
3) The performance classification of spam is further improved.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

VI. SYSTEM ARCHITECTURE

The system architecture shows the procedure followed for classification of mail into spam or not a spam(ham).

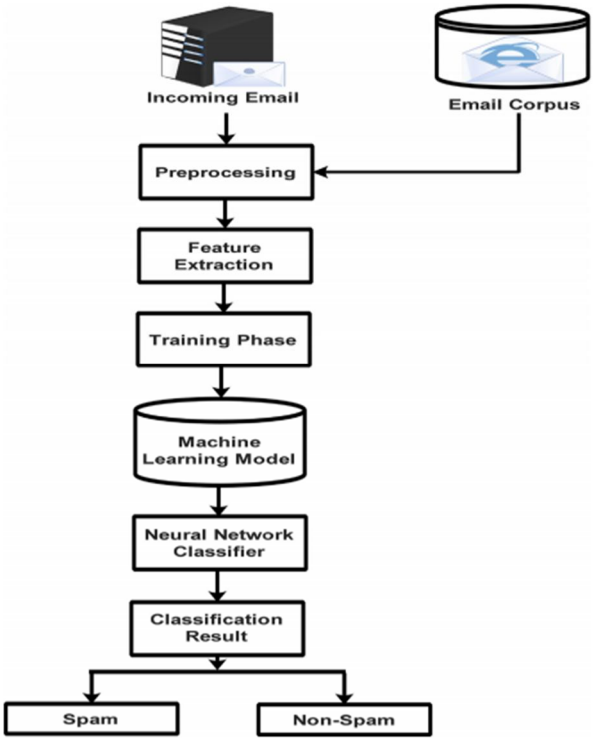


Figure2:Architecture of Classification of Spam mail

The architecture includes various steps such as uploading dataset, preprocessing the data i.e., splitting the data into training set and testing set, training the model by applying the respective classification algorithm and finally classifying whether the given mail is spam or not.[1]

VII. MODULES

A. Upload SpamBase Dataset

The selecting and uploading 'spambase.data' dataset and then click on 'Open' button to load dataset. Then the dataset is loaded.[1]

B. Preprocess Dataset

Preprocessing is the second module in our project. To read all values from dataset and then split data into train and test part where application used 80% dataset for training and 20% dataset for testing.[1]

C. Run KNN, Naive Bayes & Multilayer Perceptron Algorithms

We have to run all 3 algorithms and get their prediction metrics, we get evaluation metrics such as accuracy, recall and precision for all 3 algorithms.[1]

KNN Precision : 82.20557161921708
KNN Recall : 78.73597150143385
KNN Accuracy : 81.65038002171553
Naive Bayes Precision : 89.81616472279917
Naive Bayes Recall : 89.56607162348553
Naive Bayes Accuracy : 90.22801302931596
MLP Precision : 92.84925576417939
MLP Recall : 93.70516089980849
MLP Accuracy : 93.48534201954396

Figure3: Evaluation metrics for KNN, Naïve Bayes and MLP Algorithm

D. Run SVM, Decision Tree & AdaBoost Algorithms

First we have to run Run SVM[13], Decision Tree & AdaBoost Algorithms. Then we will get metrics for SVM, decision tree and AdaBoost algorithms.[1]

SVM Precision : 74.61791831357048
SVM Recall : 69.46798376613712
SVM Accuracy : 74.0499457111835
Decision Tree Precision : 91.77983640484872
Decision Tree Recall : 92.06713833513601
Decision Tree Accuracy : 92.29098805646036
AdaBoost Precision : 94.55727354081735
AdaBoost Recall : 94.46575111384543
AdaBoost Accuracy : 94.78827361563518

Figure4: Evaluation Metrics for SVM, Decision Tree and AdaBoost Algorithm

E. Run Random Forest & CNN Algorithm

We should run Random Forest & CNN Algorithm, then we get evaluation metrics for CNN and Random Forest algorithms.[1]

Random Forest Precision : 74.61791831357048
Random Forest Recall : 69.46798376613712
Random Forest Accuracy : 74.0499457111835
CNN Precision : 90.93372584541063
CNN Recall : 90.79331407594631
CNN Accuracy : 86.80721521377563

Figure5: Evaluation Metrics for Random Forest and CNN Algorithm

F. Accuracy Comparison Graph

In graph x-axis represents algorithm name and y-axis represents accuracy of all those algorithms and from above graph we can conclude that MLP neural network give better prediction accuracy compare to all other algorithms.[1]

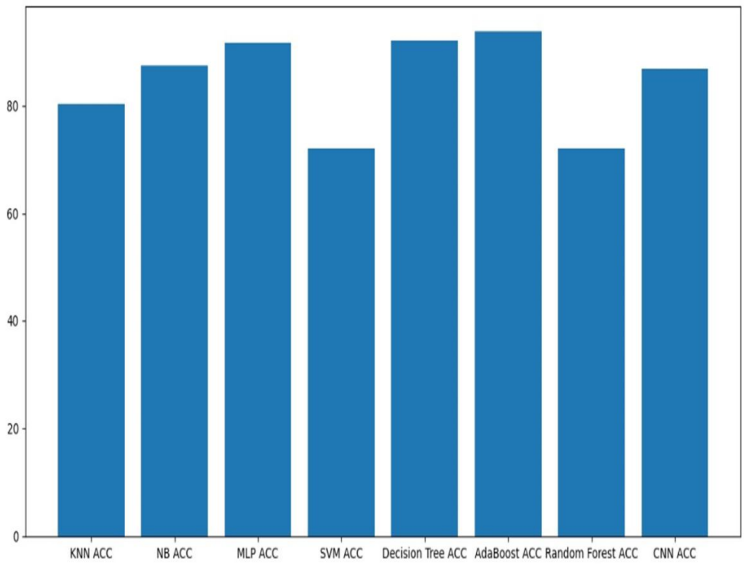


Figure6: Accuracy Comparison Graph

G. Recall Comparison Graph

In graph x-axis represents algorithm name and y-axis represents Recall values of all those algorithms.[1]

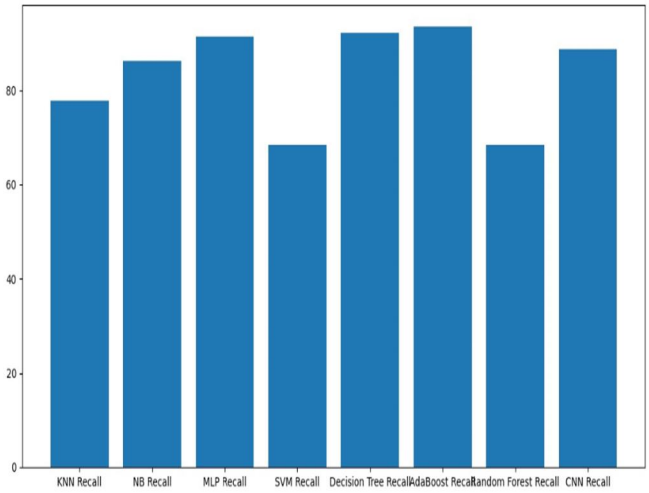


Figure7: Recall Comparison Graph

H. Precision Comparison Graph

In graph x-axis represents algorithm name and y-axis represents Precision values of all those algorithms.[1]

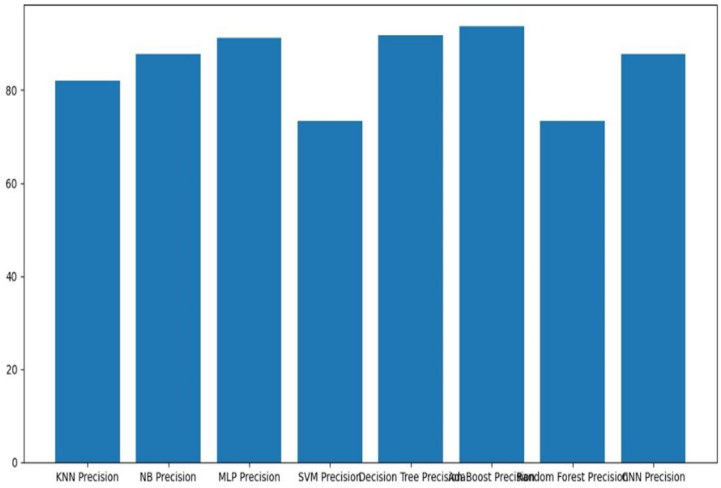


Figure8: Precision Comparison Graph

VIII. TECHNOLOGY USED IN PROJECT

A. User Interface

The user interface of this system is a user friendly python Graphical User Interface created using tkinter module of python.

B. Hardware Interfaces

The interaction between the user and the console is achieved through python capabilities.

C. Software Interfaces

The required software is python.

D. Operating Environment

Windows 10

1) Hardware Requirements

a)	Processor	-	Pentium –IV
b)	Speed	-	1.1 Ghz
c)	RAM	-	256 MB(min)
d)	Hard Disk	-	6 GB and above
e)	Key Board	-	Standard Windows Keyboard
f)	Mouse	-	Two or Three Button Mouse
g)	Monitor	-	SVGA

2) Software Requirements

h)	Operating System	-	Windows 10
i)	Programming Language	-	Python 3.8(Pycharm Editor)

IX. RESULT

In this review, we compared various machine learning algorithms such as SVM (Support Vector Machine), Random Forest, Decision Tree, CNN (Convolutional Neural Network), KNN(K Nearest Neighbor), MLP(Multi-Layer Perceptron), Adaboost (Adaptive Boosting) ,Naïve Bayes algorithm to predict or classify into spam emails**.** There are various SPAM datasets such as SPAM ARCHIVE, SPAMBASE, LINGSPAM etc to perform this experiment. We used SPAMBASE dataset to evaluate performance of above algorithms in terms of Accuracy,Precision and Recall.

In all if we compare the results by deploying the modules and by observing the results through the three graphs depicted above, we observe that MLP, Decision Tree and AdaBoost Algorithm give better accuracy, precision and recall values when we compare with other algorithms.

X. CONCLUSION

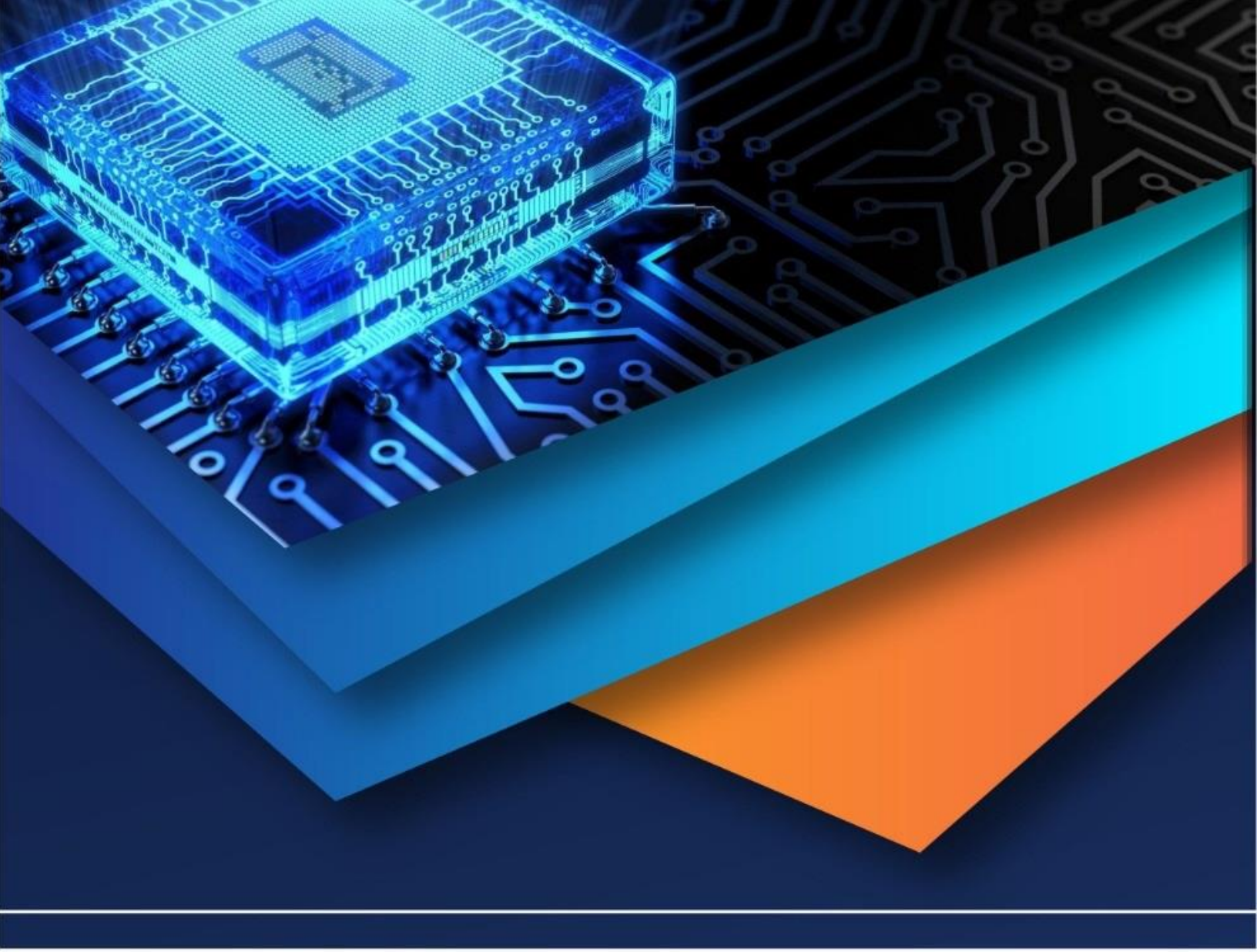
In the study, we analyzed machine learning techniques and their application to the field of spam filtering. A review of the algorithms been applied for classification of messages as either spam or ham is provided. The system architecture of email spam filter and the processes involved in filtering spam emails were looked into. The paper surveyed some of the publicly available datasets and performance metrics that can be used to measure the effectiveness and efficiency of any spam filter. The challenges of the machine learning algorithms in efficiently handling the menace of spam was pointed out and comparative studies of the machine learning techniques available in literature was done.

XI. ACKNOWLEDGEMENT

The authors would like to acknowledge the support of the Chairman, Director, Head of the Department, Department of Computer Science and Engineering, and project guides of CMR Technical Campus, Medchal, Hyderabad, Telangana, for their encouragement to the authors.

REFERENCES

[1] https://github.com/saiteja0268/IVCSED-CMRTC-SpamFilter-MajorProject-Team1-
[2] AK Sharma, S Sahni - International Journal on Computer Science and Enginnering…, 2011 – Citeseer
[3] https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9
[4] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering", *Artif. Intell. Rev.*, vol. 29, pp. 63-92, Sep. 2008
[5] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering", *Expert Syst. Appl.*, vol. 36, pp. 10206-10222, Oct. 2009.
[6] Abdul Manaan Dar, Ankit Kumar, Abhinav Raj, Jay Prakash Kumar, Monika P ,"A LITERATURE SURVEY ON FILTERING EMAILS".
[7] EmmanuelGbengaDada[a], Joseph StephenBassi[a], HarunaChiroma[b], Shafi'i MuhammadAbdulhamid[c], Adebayo OlusolaAdetunmbi[d], Opeyemi EmmanuelAjibuwa[e]: "Machine learning for email spam filtering: review,approaches and open research problems",Heliyon, Volume 5, Issue 6, June 2019.
[8] Harjot Kaur ,Er.Prince Verma:International Journal of Engineering Sciences and Research Technology Survey on "Email Spam Detection using Supervised Approach with Feature Selection" : April,2017
[9] International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 12, December 2014 Copyright to IJARCCE DOI 10.17148/IJARCCE 8688 " A survey on spam detection techniques"- Anjali Sharma , Manisha , Dr.Manisha , Dr.Rekha Jain.
[10] https://archive.ics.uci.edu/ml/datasets/Spambase
[11] "Analysis of Naïve Bayes Algorithm for Email Spam Filtering across MultipleDatasets": Nurul Fitriah Rusland, Norfaradilla Wahid, Shahreen Kasim and Hanayanti, Volume 226, International Research and Innovation Summit (IRIS2017) 6–7 May 2017, Melaka,Malaysia.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY