Random Forest Technique for E-mail Classification

Bhagyashri U. Gaikwad[1], P. P. Halkarnikar[2]
[1]*M. Tech Student, Computer Science and Technology, Department of Technology, Shivaji University, Kolhapur, Maharashtra, India,* bhagyashrigkwd@gmail.com
[2]*Associate Professor, Computer Science & Engineering, Pad. Dr. D.Y. Patil Institute of Engineering, Management and Research, Akurdi, Pune, Maharashtra, India,* pp_halkarnikar@rediffmail.com

Abstract— Email has been an efficient and popular communication mechanism as the number of Internet users increase. Therefore, email management is an important and growing problem for individuals and organizations because it is prone to misuse. The blind posting of unsolicited email messages, known as spam, is an example of misuse. Spam is commonly defined as the sending of unsolicited bulk email that is, email that was not asked for by multiple recipients. The classification algorithms such as Neural Network (NN), Support Vector Machine (SVM), and Naïve Bayesian (NB) are currently used in various datasets and showing a good classification result. This paper described classification of emails by Random Forests Technique (RF). RF is ensemble learning technique. A data mining technique called “Ensemble learning” consists of methods that generate many classifiers like decision trees and aggregates the results by taking a weighted vote of their predictions is developed. First the Body of the message is evaluated and after preprocessing the tokens are extracted. Then using a term selection method, the best discriminative terms are retained and other terms are removed. Then iterative patterns are extracted and a feature vector is built for each sample. Finally Random Forest is applied as classifier. If identified category is 0 then it is non-spam otherwise if identified category is 1 then it is spam.

Index Terms— Decision Tree, Data pre-processing, Feature selection, Random Forest, Spam.

1. INTRODUCTION

E-mail has been an efficient and popular communication mechanism as the number of Internet users increase. Therefore, email management is an important and growing problem for individuals and organizations because it is prone to misuse. The blind posting of unsolicited email messages, known as spam, is an example of misuse. Spam is commonly defined as the sending of unsolicited bulk email - that is, email that was not asked for by multiple recipients. A further common definition of a spam restricts it to unsolicited commercial email, a definition that does not consider non-commercial solicitations such as political or religious pitches, even if unsolicited, as spam. Email was by far the most common form of spamming on the internet. [2]

Spammers collect e-mail addresses from chatrooms, websites, customer lists, newsgroups, and viruses which harvest users' address books, and are sold to other spammers. In recent years, spam emails lands up into a serious security threat, and act as a prime medium for phishing of sensitive information Addition to this, it also spread malicious software to various user. An average user on the internet gets about 10-50 spam emails a day and about 13 billion pieces of unsolicited commercial e-mail are sent each day, which represents about half of all e-mail sent.[1]

It was reported an American received 2200 pieces spam e-mail on average in 2002. Increasing by 2% per month, it will reach 3600 pieces spam e-mail in 2007. A survey by CNNIC found that every email user in China received 13.7 piece emails per week in 2004, including 7.9 piece spam emails. In America, spam emails cost enterprises up to 9 billions per year. [3] A study reported that spam messages constituted approximately 60% of the incoming messages to a corporate network. Without appropriate counter-measures, the situation will become worse and spam email will eventually undermine the usability of email Anti-spam legal measures are gradually being adopted in many countries. In China, some experts advocated that an effective anti-spam e-mail measure should be carried out as early as possible. In 2003, AOL, Microsoft, EarthLink and Yahoo sued hundreds of marketing companies and individuals for sending deceptive spam using a new federal law called the CAN-SPAM Act, which prohibits such activities. But these legal measures have had a very limited effect so far due to Internet’s open architecture. Hence, apart from legal measures, we should make use of some effective anti-spam e-mail technological approaches too. At present, most anti-spam e-mail approaches, which are too simple to stop spam e-mail efficiently, block spam messages by blacklist of frequent spammers. [5]

With the proliferation of direct marketers on the Internet and the increased availability of enormous Email address mailing lists, the volume of junk mail (often referred to colloquially as spam") has grown tremendously in the past few years. As a result, many readers of E-mail must now spend a non-trivial portion of their time on-line wading through such unwanted messages. Moreover, since some of these messages can contain offensive material (such as graphic pornography), there is often a higher cost to users of actually viewing this mail than simply the time to sort out the junk. Lastly, junk mail not only wastes user time, but can also quickly fill-up file server storage space, especially at large sites with thousands of users who may all be getting duplicate copies of the same junk mail. As a result of this growing problem, automated methods for filtering such junk from legitimate E-mail are becoming necessary.[4] Automatic email spam classification contains more challenges because of unstructured information, more number of features and large number of documents. As the usage increases all of these features may adversely affect performance in terms of quality and speed. Many recent algorithms use only relevant features for classification.

Text classification including email classification presents challenges because of large and various number of features in the dataset and large number of documents. Applicability in these datasets with existing classification techniques was limited because the large number of features makes most documents undistinguishable. The classification algorithms such as Neural Network (NN), Support Vector Machine (SVM), and Naïve Bayesian (NB) are currently used in various datasets and showing a good classification result. [2]

This paper described classification of emails by Random Forests Technique (RF). RF is ensemble learning technique. A data mining technique called “Ensemble learning” consists of methods that generate many classifiers like decision trees and aggregates the results by taking a weighted vote of their predictions is developed. First the Body of the message is evaluated and after preprocessing the tokens are extracted. Then using a term selection method, the best discriminative terms are retained and other terms are removed. Then iterative patterns are extracted and a feature vector is built for each sample. Finally Random Forest is applied as classifier. If identified category is 0 then it is non-spam otherwise if identified category is 1 then it is spam.

Outline of this paper:

Section 2 presents related works on email spam classification, Section 3 presents framework of the proposed system, Section 4 presents Implementation of Random Forest, Section 5 gives result & analysis. Finally Section 6 presents conclusion and future work.

2. RELATED WORK

Vikas P. Deshpande, Robert F. Erbacher, proposed An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques in which efficient anti-spam filter that would block all spam, without blocking any legitimate messages is a growing need. To address this problem, they examine the effectiveness of statistically-based approaches Naïve Bayesian anti-spam filters, as it is content-based and self-learning (adaptive) in nature. Additionally, they designed a derivative filter based on relative numbers of tokens. They train the filters using a large corpus of legitimate messages and spam and also test the filter using new incoming personal messages.[6]Ahmed Obied proposed Bayesian Spam Filtering in which he describes a machine learning approach based on Bayesian analysis to filter spam. The filter learns how spam and non spam messages look like, and is capable of making a binary classification decision (spam or non-spam) whenever a new email message is presented to it. The evaluation of the filter showed its ability to make decisions with high accuracy. [7]

In addressing the growing problem of junk E-mail on the Internet, Mehran Sahami & Susan Dumaisy examine methods for the automated construction of filters to eliminate such unwanted messages from a user's mail stream. By casting this problem in a decision theoretic framework, they are able to make use of probabilistic learning methods in conjunction with a notion of differential misclassification cost to produce filters. In order to build probabilistic classifiers to detect junk E-mail, they employ the formalism of Bayesian networks.[4] Denil Vira, Pradeep Raja & Shidharth Gada present An Approach to Email Classification Using Bayesian Theorem. They propose an algorithm for email classification based on Bayesian theorem. The purpose is to automatically classify mails into predefined categories. The algorithm assigns an incoming mail to its appropriate category by checking its textual contents. The experimental results depict that the proposed algorithm is reasonable and effective method for email classification. [8]

Raju Shrestha and Yaping Lin present the new approach to statistical Bayesian filter based on co-weighted multi area information. This new algorithm co-relates the area wise token probability estimations using weight coefficients, which are computed according to the number of occurrences of the token in those areas. Experimental results showed significant improvement in the performance of spam filtering than using individual area-wise as well as using separate estimations for all areas.[17] Michal Prilepok1, Jan Plato proposed Bayesian Spam Filtering with NCD in which a novel variant of

Classic Bayesian filter with combination of Normaliced Compressed Distance was described. This combined filter was tested as filter for spam identification. In addition to Classical implementation of Bayesian filter, two versions of combination with NCD were implemented. The first version uses NCD for all emails which have spamcity higher than 0.5. The second version uses NCD only, when the spamcity was in the interval from 0.5 to 0.75. The second version is much faster than the first version and its speed is almost the same as speed of Classical Bayesian filter. Both new developed versions have worse efficiency in successful marking of non spam emails. The overall efficiency of both new algorithms was better than the original filter.[18]Georgios Paliouras & Vangelis Karkaletsis present Learning to Filter Spam E-Mail a Comparison of a Naïve Bayesian and a Memory-Based Approach in which they investigate the performance of two machine learning algorithms in the context of anti-spam Filtering. They investigate thoroughly the performance of the Naive Bayesian filter on a publicly available corpus, contributing towards standard benchmarks. At the same time, we compare the performance of the Naive Bayesian filter to an alternative memory based learning approach, after introducing suitable cost-sensitive evaluation measures. Both methods achieve very accurate spam filtering, outperforming clearly the keyword-based filter of a widely used e-mail reader. [9] Zhan Chuan, LU Xian-liang proposed An Improved Bayesian with Application to Anti-Spam Email in which they presents a new improved Bayesian-based anti-spam e-mail filter. They adopt a way of attribute selection based on word entropy, use vector weights which are represented by word frequency, and deduce its corresponding formula. It is proved that their filter improves total performances apparently. [5] Qiang WANG & Xinming MA proposed an ensemble learning and decision tree based approach. In this, a novel classification method based decision tree and ensemble learning is introduced to classify the spam email effectively. An experimental evaluation of different methods is carried out on a public spam email dataset. The experimental results suggest that the proposed method generally outperforms benchmark techniques.[10]

Prajakta Ozarkar, & Dr. Manasi Patwardhan made use of Random Forest and Partial Decision Trees algorithm to classify spam vs non-spam emails. These algorithms outperformed the previously implemented algorithms in terms of accuracy and time complexity. As a pre-processing step they have used feature selection methods such as Chi-square, Information gain, Gain ratio, Symmetrical uncertainty, Relief, OneR and Correlation. This allowed to select subset of relevant, non redundant and most contributing features to have an added benefit in terms of improvisation in accuracy and reduced time complexity. [12] Dario Nappa & Saeed Abu-Nimeh investigated the predictive accuracy of six classifiers on a phishing data set. The classifiers included Logistic Regression (LR), Classification and Regression Trees (CART), Bayesian Additive Regression Trees (BART), Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NNet). They constructed a data set from 1171 raw phishing emails and 1718 legitimate emails, where 43 features were trained and tested to predict phishing emails. During training and testing they used 10-fold cross-validation and averaged the estimates of all 10 folds (sub-samples) to evaluate the mean error rate for all classifiers. The results showed that, when legitimate and phishing emails are weighted equally, RF outperforms all other classifiers with an error rate of 07.72%, followed by CART, LR, BART, SVM, and NNet respectively. NNet achieved the worst error rate of 10.73%. Although RF outperformed all classifiers, it achieved the worst false positive rate of 08.29%.LR had the minimum false positive rate of 4.89%.[11]

3. FRAMEWORK OF THE PROPOSED SYSTEM

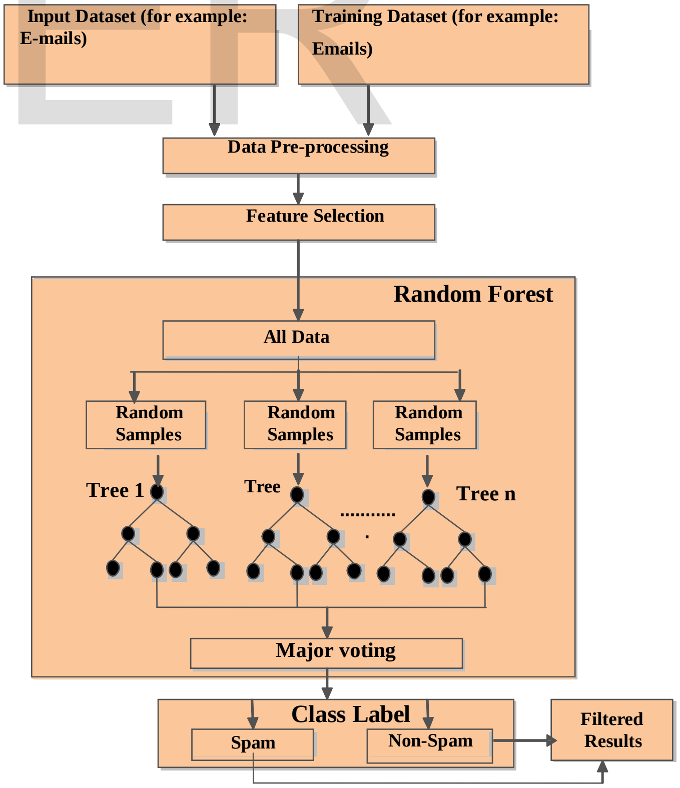


Fig.1 Proposed System Architecture

The overall design of the proposed system is given in Fig.1 Collections of emails are dataset required for training & testing purposes retrieved from following website:
http://csmining.org/index.php/spam-email-datasets-html
Proposed System consists of following steps:

1) Data pre-processing:- Pre-processing is considered as an important step in text mining. There are three steps in preprocessing task for email classification, which are tokenization, stop word removal and stemming. First step used is tokenization. In tokenizing process, all symbols (@, #, %,$), punctuations and numbers will be removed. The remaining strings will be split up into tokens. Second step is stopword removal. Many of the most frequently used words in English are useless in Information Retrieval (IR) and text mining. These words are called 'Stop words' .Stop-words, which are language-specific functional words, are frequent words that carry no information (i.e., pronouns, prepositions, conjunctions). [13] In this step, the common words, which are the most frequent words that exist in a document like ‘we’, ‘are’, ‘is’ and etc are removed. In English language, there are about 400-500 Stop words. Stop word list is based on word frequency. This process will identified which words those match with the stop word lists by comparing both of them. Removing these words will save spaces for storing document contents and reduce time taken during the searching process. Third step is stemming, “Stemming” means finding the origin of the words and removing prefixes and postfixes. By using Stemming, forms of a word, like adjectives, nouns and, verbs, are converted to homological-like word. For instance, both ‘capturing’ and ‘captured’ are converted to a same word, ‘capture’.

2) Feature selection: - Feature selection involves analyzing data (such as a bunch of average emails) and determines which features (words) will help the most in classification, which can then be used to train a classifier. One of Feature selection method is TF. Term frequency of each word in a document (TF) is a weight which depends on the distribution of each word in documents. [14] It expresses the importance of the word in the document.

3) Random Forests Algorithm: - The Random forest is a meta-learner which consists of many individual trees. Each tree votes on an overall classification for the given set of data and the random forest algorithm chooses the individual classification with the most votes. Each decision tree is built from a random subset of the training dataset, using what is called replacement, in performing this sampling. That is, some entities will be included more than once in the sample, and others won't appear at all. In building each decision tree, a model based on a different random subset of the training dataset and a random subset of the available variables is used to choose how best to partition the dataset at each node. Each decision tree is built to its maximum size, with no pruning performed. Together, the resulting decision tree models of the Random forest represent the final ensemble model where each decision tree votes for the result and the majority wins.

4) Class Label: - Depending on index value of max value calculated for trees getting class label spam or non-spam. If category is zero then class is labelled as non-spam & if category is one then class is labelled as spam.

4. IMPLEMENTATION

Ensemble classification methods are learning algorithms that construct a set of classifiers instead of one classifier, and then classify new data points by taking a vote of their predictions is developed. Ensemble learning provides a more reliable mapping that can be obtained by combining the output of multiple classifiers. [15] Fig.2 illustrates ensemble learning. Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. The algorithm was developed by Leo Breiman and Adele Cutler in the middle of 1990th.The method combines Breiman's "bagging" idea and the random selection of features.

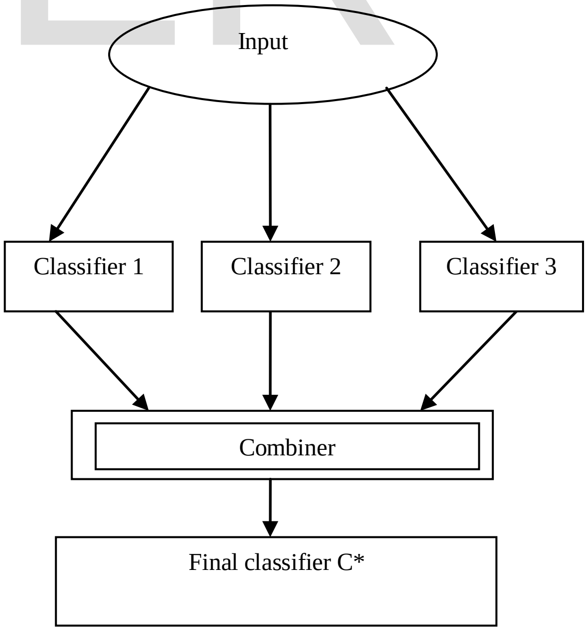
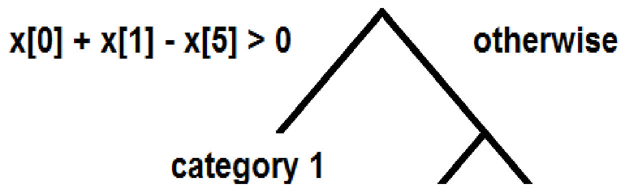


Fig.2 Ensemble Learning

Usually, the data to be categorized is known in the form of vectors and the process of categorization is navigation of each vector down along the binary tree, where special condition is applied at each node for choosing left or right direction:



The estimated category is sitting at the bottom branch of the tree and is identified when this branch is reached. The word 'forest' in the name is used because categorization is decided not by the single tree but by the large set of trees called forest. And, if trees provide different categories, the right one is selected as mode value. The word 'random' in the name is used because the nodes and switching conditions are created at a construction by using random sampling from the training set. This is also true for this particular condition shown in the above picture X [0] + X [1] - X [5] > 0. [16] It has to be created at a construction for particular group of training vectors and their components randomly but non-the-less navigating to the direction that can provide right categorization when applied for the data. Such condition is called classifier and is the most important part of the algorithm, critical for its efficiency. In each tree, the training data subset used to grow the tree is called in-of-bag (IOB) data, and the data subset formed by the remaining data is called out-of-bag (OOB) data. Since OOB data is not used to build trees, it can be used to test the OOB accuracy of each tree.

Each tree is constructed using the following algorithm:

- Let the number of training cases be N, and the number of variables in the classifier be M.
- The number of m input variables to be used to determine the decision at a node of the tree; m should be much less than M.
- Choose a training set for this tree by choosing N times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.
- For each node of the tree, randomly choose *m* variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.
- Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

Following are Steps for Classify email spam or non-spam by random forest.

Firstly Current user e-mail contents are copied into S0001A.txt file. Then when user clicked on RF label of home page then processing of RF algorithm have been performed which is as follows:

- First get data of each file in bytes & if any spaces then used character filter for removing spaces. Use word counter variable for storing position of word with word frequency.
- Creating DocWordMatrix.dat empty file & write data such as document index, position of word & frequency of word in this DocWordMatrix.dat file. DocWordMatrix.dat is document-term matrix in binary format
- Then read data from docwordmatrix.dat file & build matrix in rows & columns form. i.e. as follows:-

  m_data[row][col]=[0][{int 1558}]
  :
  :
  [95][{int 1558}]

- Next calculate random samples from training samples.
- Next start to build the forest.
- Here 96 known Categories & 49 training Samples are taken.
- For building forests required to build different trees & for building tree need to calculate node.
- Calculate node value using classifier & threshold function.
- After getting node value next step is to find left & right of that node. For calculating left & right node first need to calculate indicators using split sample function. i.e. as follows:

  indicator = [0] 0
  [1] 1
  :
  :
  [29]1

Using these indicators we find left & right vectors. i.e. suppose b is variable that indicates value of indicator then

if (b>0) then left vector

& right vector is length of indicators minus total left vectors.

- After getting left & right samples again calculate node value using classifier & threshold function. In this way split node until single sample remaining. Get category of that Remaining sample from known categories. After splitting all nodes one tree is completed.
- Similarly all trees are created. Here 200 trees are created. Each tree has different random samples.

After creating all 200 trees forest building process is completed.

- Next step is categorizing document as spam or non-spam. Here training samples which are not declared in training samples variable i.e. testing samples are only taken for categorization.

For example: k=1 i.e. 1 is document number & if you want to get category i.e. spam or non-spam of that document. Then cosine value is calculated between classifier of 0th tree & classifier of that document number 1& that value stored in variable f.

if(f>node.threshold)then getcateory from node.left otherwise from node.right. Similarly we get f value for all trees & get category of each tree. i.e. as follows:

```
        ++categoriescounter = [0]144
                              [1]56
```

Consider category 0 as non-spam & 1 as spam. Here 144 trees having category 0 & 56 trees having category 1. Next find max value between this value of 0 & value of 1. Here 144 is max value than 56 & 144 having index 0 so category of document 1 is 0 i.e. non-spam. So document 1 is a non-spam document.

5. RESULTS & ANALYSIS

After testing the system on datasets by Random Forests (RF) technique various performance measures such as the precision, recall & accuracy were observed. Collections of e-mails are Datasets required for training & testing by Random Forests (RF) technique retrieved from following website:
http://csmining.org/index.php/spam-email-datasets-.html

These performance measures for Random Forests technique on average of four datasets tested were calculated. Following table shows these calculated measures:-

TABLE I
PERFORMANCE MEASURES CALCULATED FOR RF

Measure	Defined as	Values (%)
Accuracy	(TP + TN) / (TP + FP + FN+TN)	92
Precision	TP / (TP + FP)	86.36
Recall	TP/ (TP + FN)	95

The Fig.3 shown below is Graph of System Tested Result of Spam or Non Spam emails for Average Accuracy (92%) of Four Datasets Tested by Random Forest (RF) Technique. These shows 40% are spam emails & 52% are non-spam emails. Also 8% are undefined that is that may be either spam or non-spam emails.

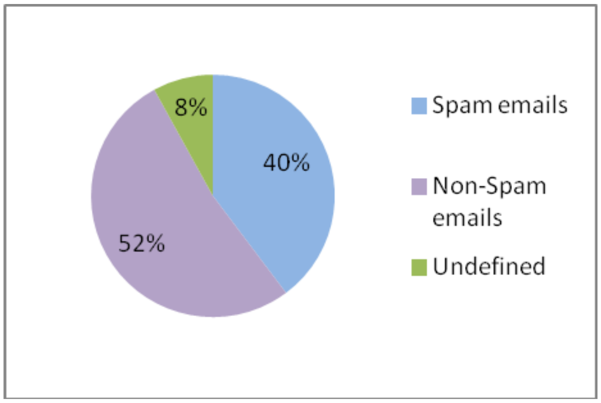


Fig. 3 Graph of System Tested Result of Spam or Non- Spam emails for Average Accuracy (92%) of Four Datasets Tested by RF Technique

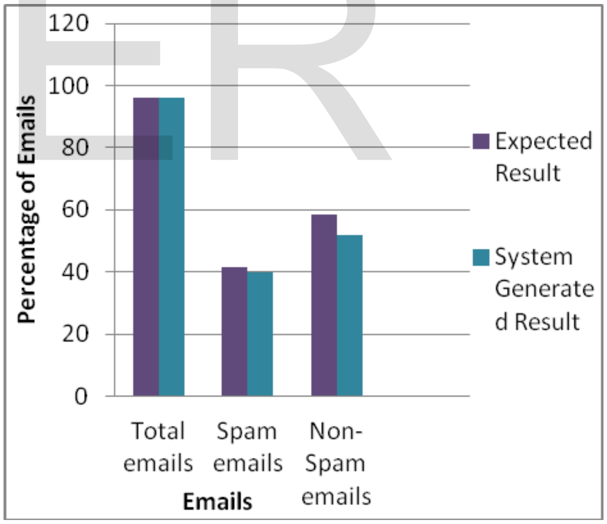


Fig. 4 Expected Result & System Generated Result of emails for Average Accuracy (92%) of Four Datasets Tested by RF Technique.

Fig.4 shown above is graph of Expected Result & System generated result of emails for average accuracy (92%) of four datasets tested by Random Forests (RF) Technique. For Expected (actual) result it shows 96% are total emails, 41.66% are spam emails & 58.33% are non-spam emails. For system generated result of emails it shows 96% are total emails, 40% are spam emails & 52% are non-spam emails.

Following Fig.5 shows graph of time shows processing time required for one dataset tested by Random Forests

Technique (RF) at a time (in seconds). Four datasets are tested by system & each dataset contains total 96 emails. Above figure shows that Dataset 1 required less processing time i.e. 10.08seconds as compare to other three datasets. Dataset 2 required 11.05seconds processing time. Dataset 3 & Dataset 4 required little more processing time than dataset 1 & 2 i.e. it required 12.01 seconds & 12.5 seconds to complete whole process.

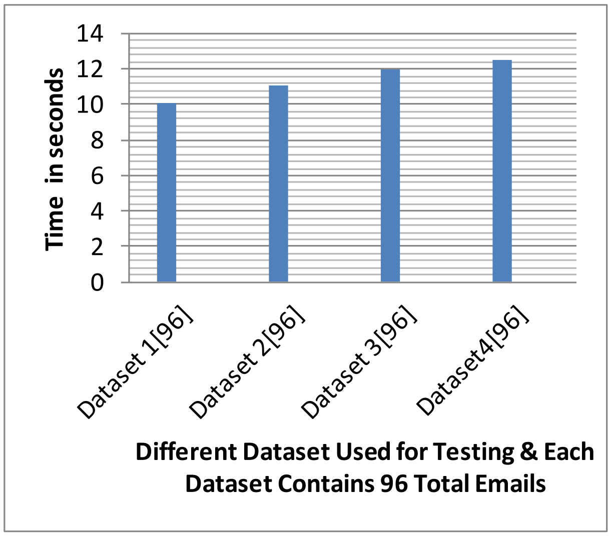


Fig. 5 Graph of time shows processing time required for One dataset tested by RF at a time (in seconds)

The obtained ROC (Receiver Operating Characteristic) curve for four datasets tested by random forests algorithm is shown in Fig. 6.

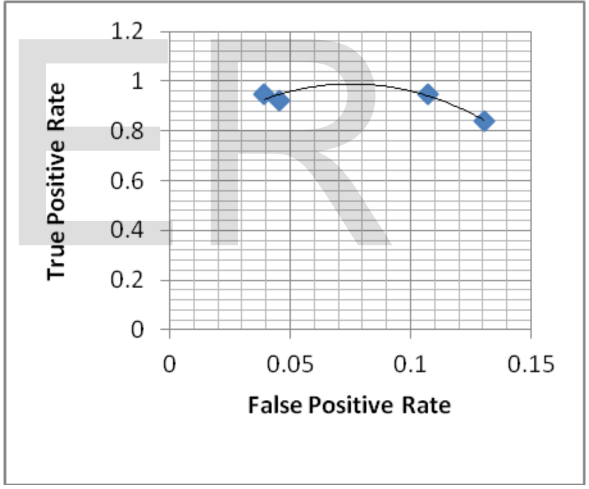


Fig. 6 ROC Graph for Four Dataset Tested by RF

The results are reported in terms of the true positive (TPR) and false positive rates (FPR). [19] Where the TPR is the number of spam messages correctly detected divided by the total number of junk e-mails.

i.e. TPR = TP / (TP + FN)

& the FPR is the number of legitimate messages misclassified as spam divided by the total number of legitimate e-mails.

i.e. FPR = FP/ (FP + TN).

Following table shows True positive rate (TPR) & False positive rate (FPR) calculated for four dataset tested by random forests (RF) technique. It is observed that higher true positive rate for dataset 2 & dataset 4 & lower false positive rate for dataset 4 as compare to other dataset.

TABLE II

TPR & FPR FOR FOUR DATASET TESTED BY RF

Dataset	True Positive Rate(TPR)	False Positive Rate(FPR)
Dataset 1	0.84	0.13043
Dataset 2	0.95	0.10714
Dataset 3	0.92	0.04545
Dataset 4	0.95	0.039215

6. CONCLUSION & FUTURE WORK

Email spam classification has received a tremendous attention by majority of the people as it helps to identify the unwanted information and threats. Therefore, most of the researchers pay attention in finding the best classifier for detecting spam emails. This paper described classification of emails by Random Forests Technique (RF). The advantage of RF is that it runs very efficiently on large datasets with high number of features, which makes it very attractive for text categorization. After testing the system Different performance measures such as the precision, recall, & the accuracy etc. were observed. The proposed system achieves 92% average accuracy of four datasets tested by Random Forests (RF) Technique.

Future work will include an adaptation of RF to deal with the problem of imbalanced classification in e-mail classification.

REFERENCES

[1] Grant Gross, *'Spam bill heads to the president'*, IDG News Service,http://www.nwfusion.com/news/2003/1209spambill.html

[2] Seongwook Youn , Dennis McLeod," *A Comparative Study for Email Classification"*, University of Southern California, Los Angeles, CA 90089 USA

[3] CNNIC. The 13th China Internet Development Status Report[R]. 2004

[4] Mehran Sahami ,Susan Dumaisy, " *A Bayesian Approach to Filtering Junk E-Mail* ",Gates Building 1A Computer Science Department Microsoft Research Stanford University Redmond, WA 98052-6399,Stanford, CA.

[5]Zhan Chuan, LU Xian-liang, ZHOU Xu, HOU Meng-shu,*"An Improved Bayesian with Application to Anti-Spam Email"*, Journal of Electronic Science and Technology of China, Mar. 2005, Vol.3 No.1

[6]Vikas P. Deshpande, Robert F. Erbacher, *"An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques"*, Proceedings of the 2007 IEEE Workshop on Information Assurance United States Military Academy, West Point, NY 20-22 June 2007.

[7] Ahmed Obied, *"Bayesian Spam Filtering"*, Department of Computer Science University of Calgary amaobied@ucalgary.ca

[8] Denil Vira, Pradeep Raja & Shidharth Gada,*"An Approach to Email Classification Using Bayesian Theorem"*, Global Journal of Computer Science and Technology Software & Data Engineering Volume 12 ,Issue 13 Version 1.0 Year 2012

[9]Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Georgios Sakkis,*"Learning to Filter Spam E-Mail: A Comparison of a Naïve Bayesian and a Memory-Based Approach"*,Software and Knowledge Engineering Laboratory Institute of Informatics and TelecommunicationsNational Centre for Scientific Research "Demokritos"153 10 Ag. Paraskevi, Athens, Greece.

[10] Qiang WANG, Xinming MA," *Spam Email Classification Using Decision Tree Ensemble"*, Journal of Computational Information Systems 8: 3 (2012) 949–956

[11] Dario Nappa, Saeed Abu-Nimeh,*"A Comparison of Machine Learning Techniques for Phishing Detection"*, SMU HACNet Lab Southern Methodist University, Dallas,TX 75275

[12] Prajakta Ozarkar, & Dr. Manasi Patwardhan,*"Efficient Spam Classification By Appropriate Feature Selection"*, International Journal of Computer Engineering and Technology (IJCET), ISSN 0976 – 6375(Online) Volume 4, Issue 3, May – June (2013)

[13] M. Basavaraju, Dr. R. Prabhakar, *" A Novel Method of Spam Mail Detection using Text Based Clustering Approach"*, Volume 5– No.4, August 2010.

[14] V.Srividhya,,R.Anitha. *"Evaluating preprocessing techniques in text categorization"*, International Journal of Computer Science & Application Issue 2010.

[15] Ratheesh Raghavan,*"Study of relationship of training set size to error rate in yet another decision tree & random forest algorithms"*,A Thesis in Computer Science, May, 2006

[16] http://www.semanticsearchart.com

[17] Raju Shrestha and Yaping Lin,"*Improved Bayesian Spam Filtering Based on Co-weighted Multi-area Information* ",Department of Computer and Communication, Hunan University,Changsha 410082, P.R. China

[18] Michal Prilepok1, Jan Platos, Vaclav Snasel, and Eyas El-Qawasmeh,"*The Bayesian Spam Filter with NCD*", Department of Computer Science, FEI, VSB - Technical University of Ostrava, 17. listopadu 15, 708 33, Ostrava-Poruba, Czech Republic

[19] El-Sayed M.,*"Learning Methods For Spam Filtering"*, College of Computer Sciences and Engineering, King Fahd University of Petroleum and Minerals, Saudi Arabia, ISBN: 978-1-61122-759-8