# Analysis of Random Forest and Naïve Bayes for Spam Mail using Feature Selection Categorization

Ms. Rachana Mishra
Oriental College of Technology, Bhopal
R. S. Thakur, Ph.D
M.A.N.I.T, Bhopal (M.P.)
(Department of computer application)

## ABSTRACT
Today, internet users are increases Spam mail is the major problem and big challenges for researcher to reduce it .Spam is commonly defined as unsolicited email messages and the goal of spam categorization is to distinguish between spam and legitimate email messages. This paper shows classification of spam mail and solving various problems is related to web space. Many machine learning algorithm are used to classified the spam and legitimate mail. This paper identify the best classification approach using bench mark dataset .The dataset consist of 9324 records and 500 attributes used for (training and testing) to build the model. This paper can play significant role to help eliminate unsolicited commercial e-mail, viruses, Trojans, and worms, as well as frauds perpetrated electronically and other undesired and troublesome e-mail. Three machines learning supervised algorithms namely naive bayes, Random Tree and Random Forest have applied on spam mail dataset using two feature selection algorithms.

## Keywords
***spam problem, spam classification, weka***

## 1. INTRODUCTION
The Internet is gradually becoming an integral part of everyday life. Internet usage is expected to continue growing and e-mail has become a powerful tool intended for idea and information exchange, as well as for commercial and social lives. Along with the growth of the Internet and e-mail, there has been a dramatic growth in spam in recent years [1, 2, 6]. The majority of spam solutions deal with the flood of spam. However, it is amazing that despite the increasing development of anti-spam services and technologies, the number of spam messages continues to increase rapidly. The increasing volume of spam has become a serious threat not only to the

Internet, but also to society. For the business and educational environment, spam has become a security issue. Spam has gone from being annoying to being expensive and risky. The enigma is that spam is difficult to define. What is spam to one person is not necessarily spam to another. Fortunately or unfortunately, spam is here to stay and destined to increase its impact around the world. It has become an issue that can no longer be ignored; an issue that needs to be addressed in a multi-layered approach: at the source, on the network, and with the end-user. Consequently, spam filtering is able to control the problem in a variety of ways. Identification and spam removal from the e-mail delivery system allows end-users to regain a useful means of communication. Many researches on spam filtering have been centered on the more sophisticated classifier-related issues. Currently, machine learning for spam classification is an important research issue. The success of machine learning techniques in text categorization has led researchers to [1,10,5] explore learning algorithms in spam filtering.

## Why do people send spam?
Spam is the electronic equivalent of junk email. People send Spam in order to sell products and services or to promote an email scam. Some Spam is purely ideological, sent by purveyors of thought. The bulk of Spam is intended, however, to draw traffic to web sites or to sell other money making schemes. Spam on the other hand can be entirely unsuccessful, but the large number of wannabe spammers waiting in the wings ensures that we will continue to receive lots of it. Spammers go to considerable effort to thwart recipients' attempts to stop spam email. They specifically design their emails to bypass your email spam filter. This can be shown by using special characters like '@' rather than the letter 'A' in words though the spam email. In this research, we present a demonstration of Spam mail categorization through classifier models determine efficiently and accuracy of spam emails. Our aim is investigation of spam and hams using different machine learning techniques. The task of email spam classification is automatically identifying unwanted, harmful, or offensive email messages before they are delivered to a user is an important, large scale application area for machine learning methods.

## Spam Problems
**Problems Related to Costs**- Spam imposes costs on all Internet users. These costs have been increasing with the growth of the number of spam messages infiltrating the Internet daily. It is difficult to calculate the total costs of spam at the global level, though estimates suggest the costs are high.[8]

### Problems Related to Privacy
The main privacy problem is that it causes significant unwanted intrusions. In addition, the collection of e-mail addresses is frequently made without the users' knowledge, much less with a specification of the purpose and consent.

These problems are exacerbated when spam is sent indiscriminately [8].

### Problems Related to Spam Content

The content of spam messages may create a problem due to fraud and deception. Fraudulent or deceptive spam can exist in a number of forms. Spammers disguise the origin of their messages because they know their messages are being blocked or filtered and they aim to entice individuals to open their messages. A common trick that spammers use is to forge the headers of messages [8, 11]].

## 2. HOW ANTI-SPAM WORKS

The mail server classifier (Spam anti-spam) works in 2 phases. The phases are given below:

### Phase 1 – Classification

The message is subjected to scrutiny to determine if it is a spam message or not. This scrutiny does not just revolve around looking for characteristics that make it spam, it also looks for characteristics that might make it not spam (this is sometimes referred to as 'ham').

### Phase 2 - Action

Once classified the message can be rejected, modified in some way (perhaps placing [SPAM] in the subject or redirecting it to a quarantine mailbox) or it can be delivered if the tests had determined that the message was a ham one [3, 8, 9, 12].

Basic steps for classification according to our analysis, Fig -1 show the detail methods for step by step classification of spam and legitimate mails.

**Step 1 :** collect the spam and anti spam data.

**Step2**: Preprocessing the data in which we have to reduce the noise, means we reduces the unwanted field. We have to clean the data.

**Step 3:** Identify the specific bad sender, read the header body.

**Step 4:** Applying particular classification tools or techniques for specified content, text and images. We have to apply this technique as per our requirement.

**Step 5:** Store the result in the form of content based or text based or image based.

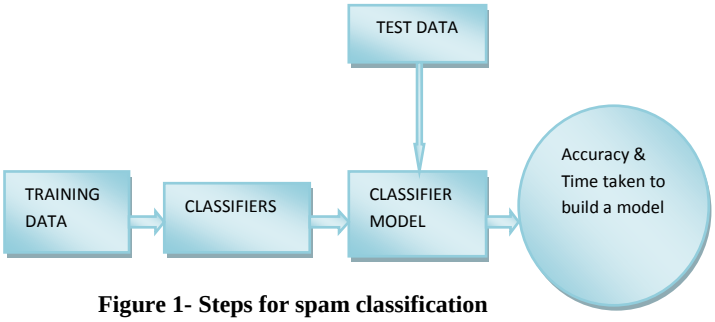


**Figure 1- Steps for spam classification**

## 3. MACHINE LEARNING PERFORMANCE CRITERIA AND ALGORITHMS

### Evaluation Measures

Accuracy and time is important factor in our dataset to build model, and comparing the algorithms. A classifier is trained to classify e-mails as non-spam and spam mail. An accuracy of 96 % may make the classifier accurate, but what if only 4-5% of the training samples are actually "spam"? clearly an accuracy of 85% may not be acceptable –the classifier could be correctly labeling only the "non-spam" samples. Instead we would like to be able to access how well the classifier can recognize "spam" samples (referred samples) how well it can recognize "non -spam" samples(referred to as negative samples).The sensitivity(recall) and specificity measures can be used ,respectively for this purpose. In addition, we may use precision to access the percentage of samples labeled as "spam" that actually are "spam" samples. The evaluation measures which are used in approach for testing process in our research work could be defined as follows:[2][3][4]

**True Positive (TP):** This states the no. of spam documents correctly classified as spam.[7]

**True Negative (TN):** This states the no. of non-spam documents correctly classified as non-spam.[7]

**False Positive (FP):** This states the no. of spam classified as non-spam.[7]

**False-Negative (FN):** This states the no. of non-spam document classified as spam.

**PRECISION** is the ratio of true positive to true and false positives. This determines how many identified objects in a class were correct.

Precision (P) = TP / (TP+FP)

**RECALL** is the ratio of true positives to the number of true positive and false negatives. This determines how many objects in a class are misclassified as something else.

Recall (R) = TP / (TP+FN)

**ACCURACY** is defined as the sum of all True positives and True Negative to the total number of test instances. This measures the overall accuracy of the classifier.

Accuracy = (TP+TN) / (TP+ TN+FP+FN)

### Confusion Matrix:

One of the methods to evaluate the performance of a classifiers using confusion matrix the number of correctly classified instances is sum of diagonals in the matrix; all others are incorrectly classified. The following terminology is often used when referring to the counts tabulated in the confusion matrix [12].

**Table-1 A Confusion Matrix**

| | | Predicted Class | |
|---|---|---|---|
| | | **Spam** | **Legitimate** |
| **Actual Class** | **Spam** | **TP** | **FN** |
| | **Legitimate** | **FP** | **TN** |

## Naïve Bays classification

A Naive Bays classifier is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bays' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model"[26].

$$\mathrm{P_r}\left[\frac{H}{S}\right] = \frac{P_r\left[\frac{S}{H}\right].P_r[H]}{P_r[H]}$$

This simple and intuitive method is based on Bayes‘s rule of conditional probability. Bayes rule says that if you have a hypothesis H and evidence E that bears on that hypothesis, then Where, Pr [H] denotes the probability of an legitimates Pr [H|S] denotes the probability of H(hams) conditional on another event S(spam).[4,10,14]

Numeric values are usually handled by assuming that they have a “normal or “Gaussian” probability distribution. For nominal attributes, we calculated counts as before, and for numeric ones we simply listed the values that occur. For the nominal attributes the counts are normalized into probabilities, whereas the mean and standard deviation for each class and each numeric attribute is to be calculated. The probability density function for a normal distribution with mean μ and standard deviation is given by the rather formidable expression:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

## Random Forest

Random Forests builds a randomized decision tree in each iteration of the bagging algorithm, and often produces excellent predictors. Random forests, introduced by Breiman, are averaged classifiers [25]. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). Each tree is grown as follows (**algorithm**) :

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.

2. 9 from spam Assassin, UseNet, Ling Spam[17] and Enron-Spam[16]. The benchmark datasets have shown in table 2.

**Table 2- Benchmark datasets**

| Name | Instance | Attributes | Work |
|---|---|---|---|
| **Usenet** | 9376 | 476 | Analysis |
| **Enron** | 7456 | 300 | Analysis |
| **SpamAssian** | 9324 | 500 | Result |

Firstly, We collect the spam dataset from easily available on spam Assassin, Eron, KDD. Datasets were in text format and transformed into Arff format.

```
@relation email_data
@attribute ac {0,1}
@attribute yes {0,1}
@attribute activities {0,1}
@attribute actually {0,1}
@attribute add {0,1}
@attribute address {0,1}
@attribute after {0,1}
@attribute go {0,1}
@attribute all {0,1}
@attribute nice {0,1}
@attribute win {0,1}
@attribute already {0,1}
@attribute also {0,1}
@attribute you {0,1}
@attribute young {0,1}
@attribute choice {0,1}
@attribute zoo {0,1}
@attribute spam,nonspam {0,1}

@data

{0,1,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,01}
```

**Fig2. ARFF format dataset**

**Table -3 (Confusion matrix for Naïve Bays)**

| Predicted Class / Actual class | Spam | Legitimate |
|---|---|---|
| **Spam** | 576 | 119 |
| **Legitimate** | 114 | 1989 |

**Table -4 (Confusion Matrix for Random Forest)**

| Predicted class / Actual class | Spam | Legitimate |
|---|---|---|
| **Spam** | **626** | **68** |
| **Legitimate** | **33** | **2070** |

**Table 5(Confusion Matrix for Random Tree)**

| Predicted class / Actual class | Spam | Legitimate |
|---|---|---|
| **Spam** | 606 | 88 |
| **Legitimate** | 37 | 2066 |

## 4. FEATURE SELECTION ALGORITHMS

Feature selection plays an important role for ML algorithms. It is not only reduces the features sets but also improve computational performance and classification accuracy. Attribute selection also known as feature selection can be used to remove model redundancy thus reducing the time required for model generation.

| Accuracy / Classifiers | Correctly classified instances | Incorrectly Classified instances | Time taken to build model |
|---|---|---|---|
| **Navies Bayes** | 91.6641% | 8.334% | 0.02sec |
| **Random Forest** | 96.389% | 3.611% | 0.5sec |
| **Random Tree** | 95.5309% | 4.4691% | 0.06Sec |

### Correlation-based Feature selection

CFS is the first of the methods that evaluate subsets of attributes rather than individual attributes [26]. At the heart of the algorithm is a subset evaluation heuristic that takes into account the usefulness of individual features for predicting the class along with

$$Consistency\ s = 1 - \frac{\sum_{i=0}^{j} |D_i||M_i|}{N}$$

Where s is an attribute subset, j is the number of distinct combinations of attribute values for s,| Di | is the number of occurrences of the $i_{th}$ attribute value combination, | Mi | is the cardinality of the majority class for the $i_{th}$ attribute value combination and N is the total number of instances in the data set.

## 5. EXPERIMENTAL RESULTS

In this section, Performance of three classification algorithms namely Navies Bays, Random Forest and Random tree has compared. The simulations were conducted using a large spam email dataset consisting of 9324 instances having 500 attributes. The UCI dataset has been modified accordingly and used mining. Table -6 shows Precision, F-measure & recall values using feature selection namely Correlation based and Consistency based algorithms. Table-7 shows classification accuracy and time taken to build model for classifiers using feature selection algorithms. Table -3 shows Confusion matrix for Naïve Bays. Table- 4 shows **c**onfusion Matrix for Random Forest. Table- 5 shows **c**onfusion Matrix for Random tree.

**Table-6 (Measures)**

| Measures / Algorithm | Precision | Recall | F-Measure | class |
|---|---|---|---|---|
| **Random Forest** | 0.95 | 0.902 | 0.925 | Spam |
| **Naive Bays** | 0.835 | 0.829 | 0.832 | spam |
| **Random Forest** | 0.968 | 0.984 | 0.976 | legitimate |
| **Naïve Bayes** | 0.944 | 0.944 | 0.945 | Legitimate |
| **Random Tree** | 0.975 | 0.965 | 0.97 | Spam |
| **Random Tree** | 0.989 | 0.992 | 0.998 | Legitimate |

**Table -7 (Accuracy results using feature selection three feature selection algorithms)**

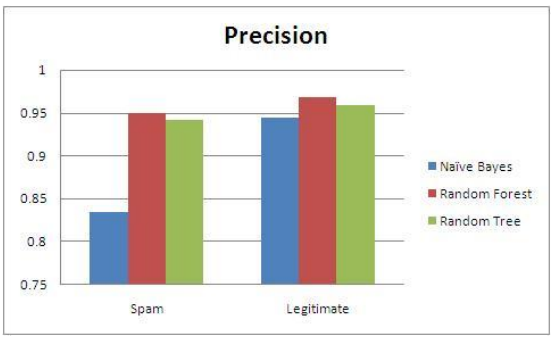


**Fig 3: Chart Graph of Correlation-Based Feature Selection Algorithm**

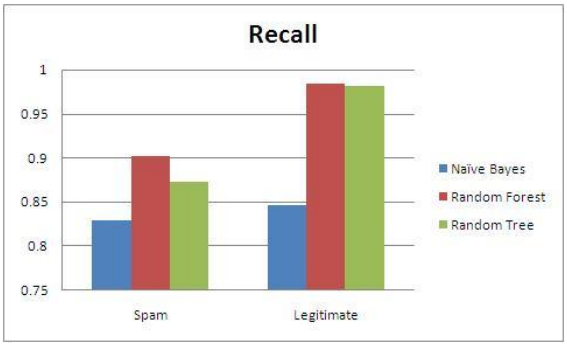


**Fig 4: Chart Graph of Correlation-Based Feature Selection**

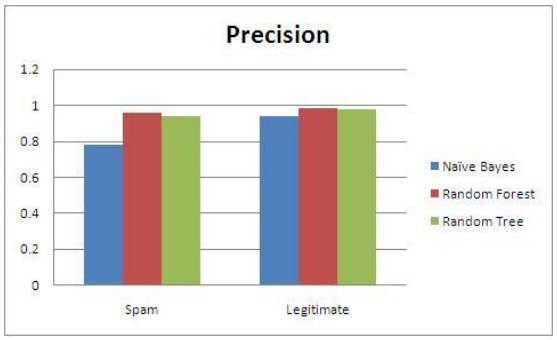


**Fig 5: Chart Graph of Consistency-Greedy Feature Selection Algorithm**

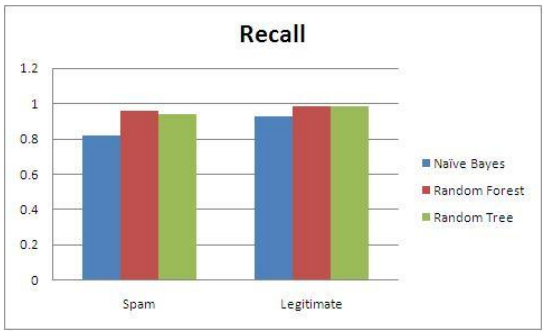


**Fig 6: Chart Graph of Consistency-Greedy Feature Selection Algorithm**

## 6. CONCLUSION AND FUTURE WORK

In this paper, spam categorization by two feature selection algorithms and benchmark dataset is used. Results analyses are comparing in terms of accuracy, precision, recall and time taken to build model using two feature selection methods. Spam measures which are helpful for identification of spam mail and legitimates. The results analysis shows best classification techniques for spam mail identification or categorization, so given measures are helpful for features selection. The conclusion from above simulation's results is that number of spam is a very serious problem which is drastically widespread. Result analysis shows that random forest classifier shows higher accuracy as compared to naives bays & Random Tree. Because accuracy of random forest in our dataset is 96.389% and navies bays accuracy is 91.6641%. Random Forest runs efficiently on large data bases. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing. Time taken is very little for Random Tree in term of accuracy. In future we will use the different classifiers to evaluate the performance. Parallel Algorithm may be developed to reduce time required for classification. Feature selection Algorithms can be used to reduce redundancy of dataset. We can also apply the other tools like Matlab, SVM, Tangra, Rapid Minor.

## 7. ACKNOWLEDGEMENT

This work is supported by research grant from MANIT, Bhopal India under Grants in Aid Scheme 2010-11, No Dean(R&C)/2010/63 dated 31/08/2010.

## 8. REFERENCE

[1]. http://www.dpw.co.santacruz.ca.us/www.santacruzcountyrecycles/Junk_Mail/index.html.

[2]. Improvising BayesNet Classifier Using Various Feature Reduction Method for Spam Classification, 1D. Shanmuga Priyaa, 2B. Kavitha, 3R. Naveen Kumar, 4K. Banuroopa 1Dept. of Information Technology, Karpagam University, India,, IJCST Vol. 1, Issue 2, December 2010.

[3]. A Novel Approach towards Image Spam Classification, M.Soranamageswari, Dr.C.MeenaInternational Journal of Computer Theory and Engineering, Vol.3, No.1, February, 2011 ,1793-8201.

[4]. Fulu Li, Mo-han Hsieh, "An empirical study ofclustering behavior of spammers and Group based Anti-spam strategies", CEAS 2006, pp 21-28, 2006.

[5]. Dhinaharan Nagamalai, Cynthia.D, Jae Kwang Lee," ANovel Mechanism to defend DDoS attacks caused by spam", International Journal of Smart Home, SERSC, Seoul, July 2007, pp 83-96.

[6]. Calton pu, Steve webb: "Observed trends in spam construction techniques: A case study of spam evolution", CEAS 2006, pp 104-112, July 27-28, 2006.

[7]. Anirudh Ramachandran, David Dagon, Nick Feamste,"Can DNS-based Blacklists keep up with Bots", CEAS 2006,CA, USA, July 27-28, 2006.

[8]. SpamCop , available at http://spamcop.net.

[9]. Internet User Forecasts by Country http:// www. etforecasts.Com.

[10]. Nigerian fraud mail Gallery http://www.potifos.com/fraud/.

[11]. Fairfax Digital http://www.smh.com.au/articles /2004/10/18.

[12]. D. Shanmuga priyaa ,b.Kavitha "Improvising Bayes Net classifier using various feature reduction method for spam classification" ,ISSN :0976-8491

[13]. Anil.K Jain, Robert P.W, Jianchang Mao "Statistical Pattern Reorganization: A Review", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 1, JANUARY 2000.

[14]. Biao Qin, Yuni Xia Sunil Prabhakar, Yicheng Tu "A Rule-Based Classification Algorithm for Uncertain Data" IEEE International Conference on Data Engineering 2009

[15]. Ziqiang Wang, Xia Sun "An Efficient Spam Filtering Algorithm Based on NPE" IEEE International Symposium on Knowledge Acquisition and Modeling Workshop,21-22Dec 2008 pp 1102 – 1104.

[16]. http://www.aueb.gr/users/ion/data/PU123ACorpora.tagz.

[17]. http://www.aueb.gr/users/ion/data/lingspam_public.tar.gz

[18]. Ravi Kiran and Indriyati Atmosukarto , "Spam or Not Spam. That is the question".

[19]. David Mertz "Spam Filtering Techniques: Comparing a Half-Dozen Approaches to Eliminating Unwanted Email" August2002Availableat:http://gnosis.cx/publish/programming/filtering-spam.html

[20]. David Mertz "Spam Filtering Techniques: Comparing a Half-Dozen Approaches to Eliminating Unwanted Email" August 2002

[21]. Available at: http://gnosis.cx/publish/ programming/filtering-spam.html

[22]. Vangelis Metsis, Ion Androutsopoulos, Georgios Paliouras,"Spam Filtering with Naive Bayes – Which Naive Bayes?" CEAS 2006 Third Conference on Email and AntiSpam July 27-28, 2006, Mountain View, California USA.

[23]. Tommi S. Jaakkola "Machine learning: lecture 7" MIT CSAIL Available at:http://www. ai.mit.edu/courses/6.867-f04/lectures/lecture-7-ho.pdf.

[24]. http://gogoshen.org/ml2005/Journal%20Paper/Journal Paper_Livingston.pdf.

[25]. http://en.wikipedia.org/wiki/Naive_bayes.

[26]. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining by Mark A Hall and Geoffrey Holmes at: http://www2.computer.org/portal/web/csdl/doi/10.1109/TKDE/2003.html