American Journal of Engineering Research (AJER)
e-ISSN : 2320-0847 p-ISSN : 2320-0936
Volume-02, Issue-10, pp-63-73
www.ajer.org

Research Paper Open Access

An Efficient Spam Filtering Techniques for Email Account

S. Roy, A. Patra, S.Sau, K.Mandal, S. Kunar
[1,2,3]*Computer Science & Technology&*[4]*Mechanical Engineering, NITTTR, Kolkata, India*
[5]*Production Engineering, Jadavpur University, Kolkata, India*

Abstract: - *Unsolicited emails, known as spam, are one of the fast growing and costly problems associated with the Internet today. Electronic mail is used daily by millions of people to communicate around the globe and is a mission-critical application for many businesses. Over the last decade, unsolicited bulk email has become a major problem for email users. An overwhelming amount of spam is flowing into user's mailboxes daily. Not only is spam frustrating for most email users, it strains the IT infrastructure of organizations and costs businesses billions of dollars in lost productivity. The necessity of effective spam filters increases. In this paper, we presented an efficient spam filter techniques to spam email based on Naive Bayes Classifier. Bayesian filtering works by evaluating the probability of different words appearing in legitimate and spam mails and then classifying them based on those probabilities.*

Keywords: - *Spam, Filters, Bayesian, Content based spam filter and Email*

I. 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 users' 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. 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 just 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 destinedto 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 [1].

In this digital age, which is the era of electronics & computers, one of the efficient & power mode of communication is the email. Undesired, unsolicited email is a nuisance for its recipients; however, it also often presents a security threat. For ex., it may contain a link to a phony website intending to capture the user's login credentials (identity theft, phishing), or a link to a website that installs malicious software (malware) on the user's computer. Installed malware can be used to capture user information, send spam, host malware, host phish, or conduct denial of service attacks as part of a "bot" net. While prevention of spam transmission would be ideal, detection allows users & email providers to address the problem today [2].

Spam filtering has become a very important issue in the last few years as unsolicited bulk e-mail imposes large problems in terms of both the amount of time spent on and the resources needed to automatically filter those messages. Email communication has come up as the most effective and popular way of communication today. People are sending and receiving many messages per day, communicating with partners and friends, exchanging files and information. E-mail data's are now becoming the dominant form of inter and intra-organizational written communication for many companies and government departments. Emails are the essential part of life now just likes mobile phones & i-pods [3].

Emails can be of spam type or non-spam type. Spam mail is also called as junk mail or unwanted mail whereas non-spam mails are genuine in nature and meant for a specific person and purpose. Information

retrieval offers the tools and algorithms to handle text documents in their data vector form. The Statistics of spam are increasing in number. At the end of 2002, as much as 40 % of all email traffic consisted of spam. In 2003, the percentage was estimated to be about 50 % of all emails. In 2006, BBC news reported 96 % of all emails to be spam.

Spam can be defined as unsolicited (unwanted, junk) email for a recipient or any email that the user do not wanted to have in his inbox. It is also defined as "Internet Spam is one or more unsolicited messages, sent or posted as a part of larger collection of messages, all having substantially identical content." There are severe problems from the spam mails, viz., wastage of network resources (bandwidth), wastage of time, damage to the PC's & laptops due to viruses & the ethical issues such as the spam emails advertising pornographic sites which are harmful to the young generations[4].

Email is the most widely used medium for communication worldwide because it's Cheap, Reliable, Fast and easily accessible. Email is also prone to spam emails because of its wide usage, cheapness & with a single click you can communicate with anyone anywhere around the globe. It hardly cost spammers to send out 1 million emails than to send 10 emails. Hence, Email Spam is one of the major problems of the today's internet, bringing financial damage to companies and annoying individual users.

- **Rule based**

Handmade rules for detection of spam made by experts (needs domain experts & constant updating of rules).

- **Customer Revolt**

Forcing companies not to publicize personal email ids given to them (hard to implement).

- **Domain filters**

Allowing mails from specific domains only (hard job of keeping track of domains that are valid for a user).

- **Blacklisting**

Blacklist filters use databases of known abusers, and also filter unknown addresses (constant updating of the data bases would be required).

- **White list Filters**

Mailer programs learn all contacts of a user and let mail from those contacts through directly (every one should first be needed to communicate his email-id to the user and only then he can send email).

- **Hiding address**

Hiding ones original address from the spammers by allowing all emails to be received at temporary email-id which is then forwarded to the original email if found valid by the user (hard job of maintaining couple of email-ids).

- **Checks on number of recipients by the email agent programs.**
- **Government actions**

Laws implemented by government against spammers (hard to implement laws).

- **Automated recognition of Spam**

Uses machine learning algorithms by first learning from the past data available (seems to be the best at current). Here, follows a brief overview of e-mail spam filtering. Among the approaches developed to stop spam, filtering is an important and popular one. It can be defined as automatic classification of messages into spam and legitimate mail. It is possible to apply the spam filtering algorithms on different phases of email transmission at routers, at destination mail server or in the destination mailbox[5]. Filtering on the destination port solves the problems caused by spam only partially, i.e., prevents end users from wasting their time on junk messages, but it does not prevent resources misuse, because all the messages are delivered nevertheless. In general, a spam filter is an application which implements a function:

f(m, L) = { cspam, if the decision is "spam" cleg, otherwise }

Where 'm' is a message or Email to be classified, L is a vector of parameters, and cspam and cleg are labels assigned to the messages. Most of the spam filters are based on a machine learning classification techniques. In a learning-based technique the vector of parameters L is the result of training the classifier on a pre collected dataset:

L = R(M),

M = {(m1, y1),…… (mn, yn)}, yi∈ { cspam, cleg}

Where m1, m2…mn are previously collected messages, y1, y2…yn are the corresponding labels, and R is the training function. In order to classify new message, a spam filter can analyze them either separately (by just checking the presence of certain words) or in groups (consider the arrival of dozen of messages with same content in five minutes than arrival of one message with the same content). In addition, learning-based filter analyzes a collection of labeled training data (pre-collected messages with reliable judgment).

This paper explores statistical learning algorithms such as Bayesian techniques for classifying spam. This spam probability and non spam probability for every word occurred in incoming message by using training sets of

words. Every word contains two frequencies one is spam frequency and other is non-spam frequency. Using this frequency calculate spam and non-spam probability [6]. If spam probability greater than non-spam probability then incoming message is considered as spam email message. After that all words have been updated in trainings set. All time training sets will be updated.

II. GENERAL CHARACTERISTICS OF SPAM

Spam is not only offensive and annoying; it causes loss of productivity, decreases bandwidth and costs companies a lot of money. Therefore, every smart company that uses email must take measures in order to block spam from entering their email systems. Although it might not be possible to block out all spam, just blocking a large proportion of it will greatly reduce its harmful effects. In order to effectively filter out spam and junk mail, the proposed system is able to distinguish spam from legitimate messages and to do this it needs to identify typical spam characteristics & practices. Once these practices are known, suitable measures can be put into place to block these messages. Of course, spammers are continually improving their spam tactics, so it is important to keep up to date on new spam practices from time to time to ensure spam is still being blocked effectively.

Spam characteristics appear in two parts of a message; email headers and message content:

2.1 Email Header

Email headers show the route an email has taken in order to arrive at its destination. They also contain other information about the email, such as the sender and recipient, the message ID, date and time of transmission, subject and several other email characteristics. Most spammers try to hide their identity by forging email headers or by relaying mail to hide the real source of the message. Since they need to send mails to a large number of recipients, spammers use certain methods for mass mailing that can be classified as pure spam practices and can therefore be identified in the email headers. Although newsletters and legitimate mailings are also sent to a large number of recipients, these will generally not display the same characteristics since the message source does not need to be concealed.

Typical email header characteristics in spam messages:

Recipient's email address is not in the To: or Cc: fields
Empty To: field
To: field contains invalid email address
Missing To: field
From:
Missing From: field
Missing or malformed Message ID
More than 10 recipients in To: and/or Cc: fields
X-mailer field contains name of popular spam ware
Bcc: header exists
X-Distribution = bulk
X-UIDL header exists
Code and space sequence exists
Illegal HTML exists

Table 1. Statistics of spam based on spam characteristics

Spam Characteristics	% of Searched mails
Recipient address not in To: or Cc: field	64%
To: field missing	34%
To: field contains invalid email address	20%
No message ID	20%
Suspect message ID	20%
Cc: filed contains more than 15 recipients	17%
From: is same as the To: field	6%
Cc: field contains more between 5-15 recipients	3%
To: field contains more between 5-15 recipients	2%
Cc: field contains more than 5-15 recipients	1%
Bcc: field exists	0%
To: field is empty	0%
From: is blank or missing	0%

2.2 Message contents

Apart from headers, spammers tend to use certain language in their emails that companies can use to distinguish spam messages from others. Typical words are free, limited offer, click here, act now, risk free, lose weight, and earn money, get rich, and (over) use of exclamation marks and capitals in the text. Spam canbe blocked by checking for words in the email body and subject, but it is important that you filter words accurately since otherwise you might be blocking legitimate mails as well.

III. DEVELOPMENT OF THE PROPOSED SYSTEM

3.1 Overview of Design Methodology

The proposed system using naive bayes classifier to classify an email is spam or not. Proposed system should be good rate for false positive and false negative. While false positive means a good email can be identified as spam email. False negative means a spam emails identified by a good email.

3.1.1 Bayes Theorem

The probability of an event may depend on the occurrence or non-occurrence of another event. This dependency is written in terms of **conditional probability**:

P(A|B) = P(A∩B) / P(B)

P(B|A) = P(A∩B) / P(A)

P(A∩B) =P(B|A) P(A) = P(A|B) P(B)

An event A is INDEPENDENT from event B if the conditional probability is the same as the **marginal probability**.

P(B|A) = P(B)

P(A|B) = P(A)

From the formulas the Bayes Theorem States the Prior probability: Unconditional probabilities of our hypothesis before we get any data or any NEW evidence. Simply speaking, it is the state of our knowledge before the data is observed. Also stated is the posterior probability: A conditional probability about our hypothesis (our state of knowledge) after we revised based on the new data.

Likelihood is the conditional probability based on our observation data given that our hypothesis holds.

P(A|B) = P(B|A) P(A)/ P(B)

P(B|A) = P(B|A) P(B)/ P(A)

Where P (A|B) is the posterior probability, P(B|A) is the likelihood and P(A) prior probability.

Thomas Bayes (c. 1702 – 17 April 1761) was a British mathematician and Presbyterian minister, known for having formulated a specific case of the theorem that bears his name: Bayes' theorem, which was published posthumously.

The following are the mathematical formalisms, and the example on a spam filter, but keep in mind the basic idea.

The Bayesian classifier uses the Bayes theorem, which says:

P(cj | d) = P(d |cj) P(cj) / P(d)

Considering each attribute and class label as a random variable and given a record with attributes (A1,A2,…, An), the goal is to predict class C. Specifically, we want to find the value of C that maximizes P(C| A1,A2,…An).

The approach taken is to compute the posterior probability P(C| A1,A2,…An) for all values of C using the Bayes theorem.

P(C | A1 A2 ...An) = P(A1 A2 ....An | C) P(C) / P(A1 A2 ....An)

So you choose the value of C that maximizes P(C| A1,A2,…An). This is equivalent to choosing the value of C that maximizes P(A1,A2,…An | C) P(C).

3.2 Evaluation Process

Naïve Bayesian prediction requires each conditional probability be non zero. Otherwise, the predicted probability will be zero.

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

In order to overcome this, we use probability estimation from one of the following:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m - estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes
p: prior probability
m: parameter

In order to classify and predict a spam email from a non spam one, the following techniques and assumptions are used:

1. Sorting according to language (spam or non spam), then words, and then count.
2. If a word does not exist, consider to approximate P(word|class) using Laplacian.
3. A learning dataset for analysis.
4. The Learning Datasetcontains each word that content filtering uses to determine if a message is spam. Beside each word, there are two numbers. The first number is the number of times that the word has occurred in non-spam e-mail messages. The second number is the number of times that the word has occurred in spam e-mail messages.

Table 2.Example of learning dataset containing words

Word	Occurred in Non Spam	Occurred in Spam
Specializing	391	4022
Graciously	2095	380
Bringing	2772	11854
Mbps	425	823
Tantra	96	52

3.3 Algorithm of the Proposed System

Input: keyword list, stop word list, ignore list, email message

Step1:- Let m be an email message. Take two variables Nonspam percent and Spam percent initialized to 1. Convert all words to lower case.

Step2:- Find out all special character from m and remove all special character from m.

Step 3:- For each word w_i in m
If w_i found in stopword list
Then w is removed from m.
End If.
End For.

Step 4:-For each word w_j in m
If w_j found in learning keyword dataset
Take spam value and calculate the probability using spam value divided by total spam value and multiply this probability with Spam percent
Else Take spam value as 1(using Laplace) and calculate the Probability using 1 divided by total spam value and Multiply this probability with Spam percent. And add to keyword dataset.
End For.

Step 4:-For each word w_k in m
If w_k found in learning keyword dataset
Take non spam value and calculate the probability using non spam value divided by total non spam value and multiply this probability with NonSpam percent
Else take non spam value as 1(using Laplace) and calculate the Probability using 1 divided by total non spam value and Multiply this probability with NonSpam percent. And add to keyword dataset.
End For.

Step 5:- Find out spam and non spam probability from learning dataset and multiply spam probability with Spam percent to get all word spampercent and similarly multiply non spam probability with non spam percent to get all word non spam percent

Step 6:- If Spam percent >Nonspam percent
Then m will be identified as Spam Email.

Else m will be legitimate Email.

3.4 Explanation of Algorithm

Input of the above algorithm is ignoring list, stop word list, and keyword list. Ignore list contains set of special character like (~,!,@,#,$,%,^,&,*,<,>,.,? etc)

First to remove from email content this ignores character. Then remove stop word from email content. Stop word list like (am, is, are etc). Rest of the word contains in email is termed as keyword. To maintain a huge learning data set using this keyword. Learning datasets contain lots of word. And each word has two properties. One spam count and another is non spam count [7]. Where spam count is no of spam email contains this word. Non spam count is the no of non spam email that contains this word.

First find out probability of mail type from the training dataset. Probability of spam email will be

P (SpamMailType) = (Total spam count from learning set/ Total count)

Similarly probability for non spam email will be

P (NonSpamMailType) = (Total spam count from learning set/ Total count)

Then take two variable example spampercent and nonspampercent and initialized to 1. Then look for percent of word in non spam. Similarly look for percent of word in spam. Final probability by multiplying each word to the total probability. Compare spampercent with nonspampercent. If spam percent greater than nonspampercent, then email is considered as spam email.

IV. IMPLEMENTATION

To implement a system, there are some software and hardware required. Some softwares as well as hardwaresare used to implement the above proposed system.

Hardware Used: 1. Windows 7

Software Used: 1.Visual studio 2010
2. Sql Server 2008
3. IIS 7.0

Some of this system snapshot is given below.

Login page of this email system

Fig 4.1 Login page

After login successful page

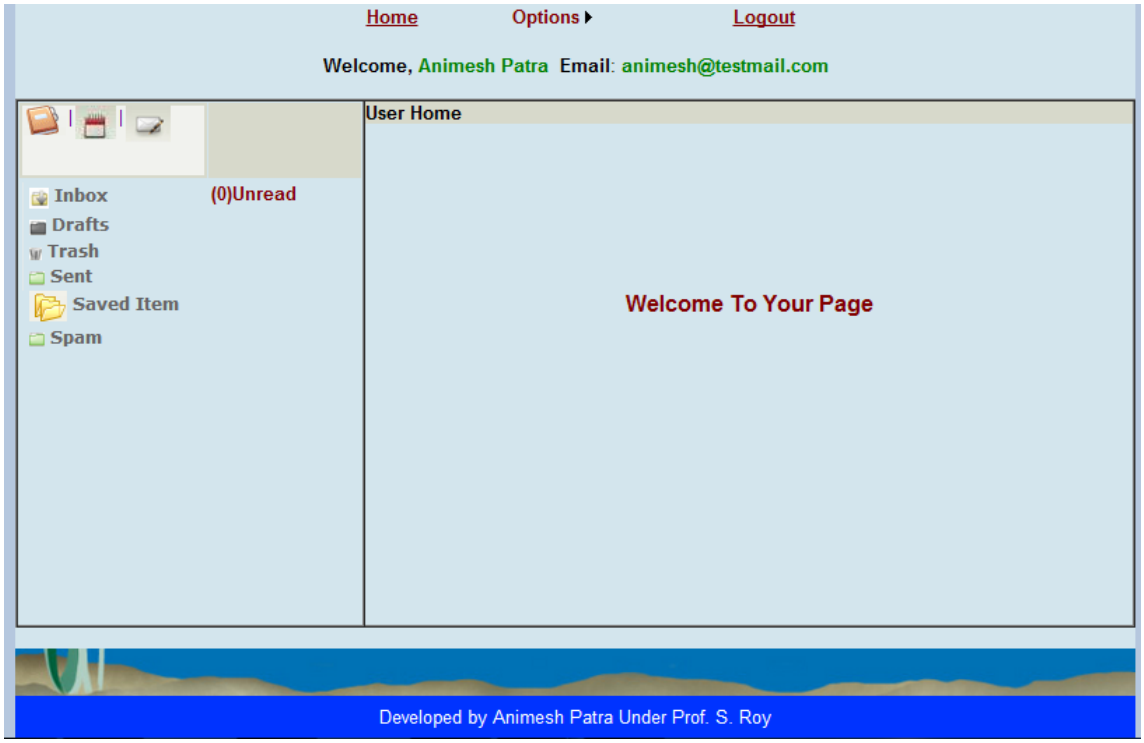

Fig 4.2. Home page

Compose mail

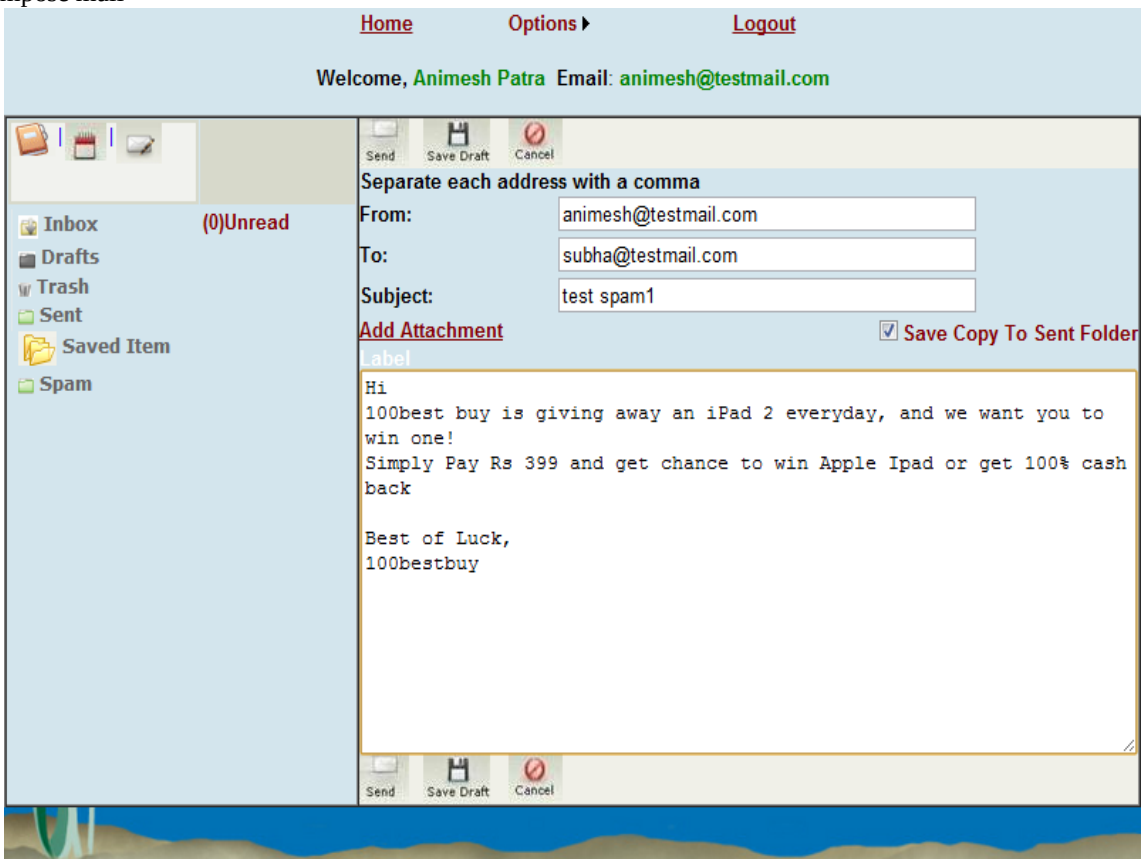

Fig 4.3 Compose mail page

After sending mail

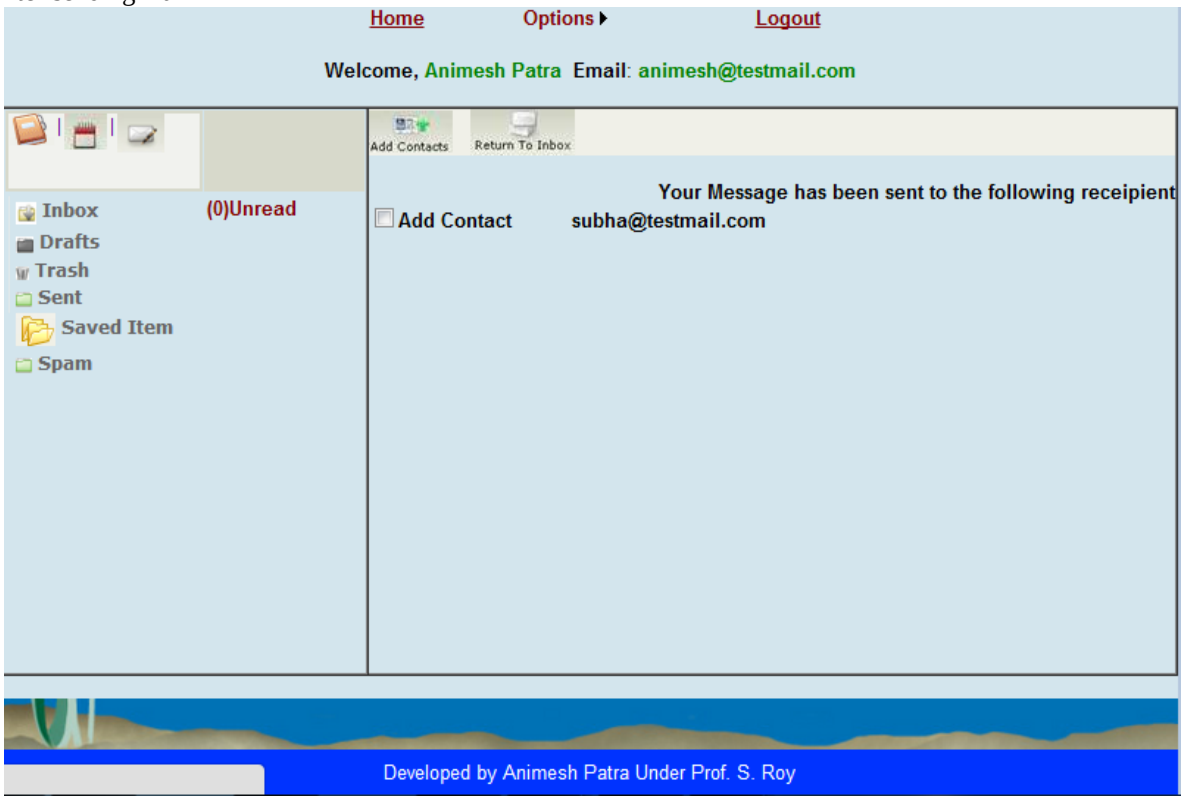

Fig.4.4 After email sending Page

After login successful page

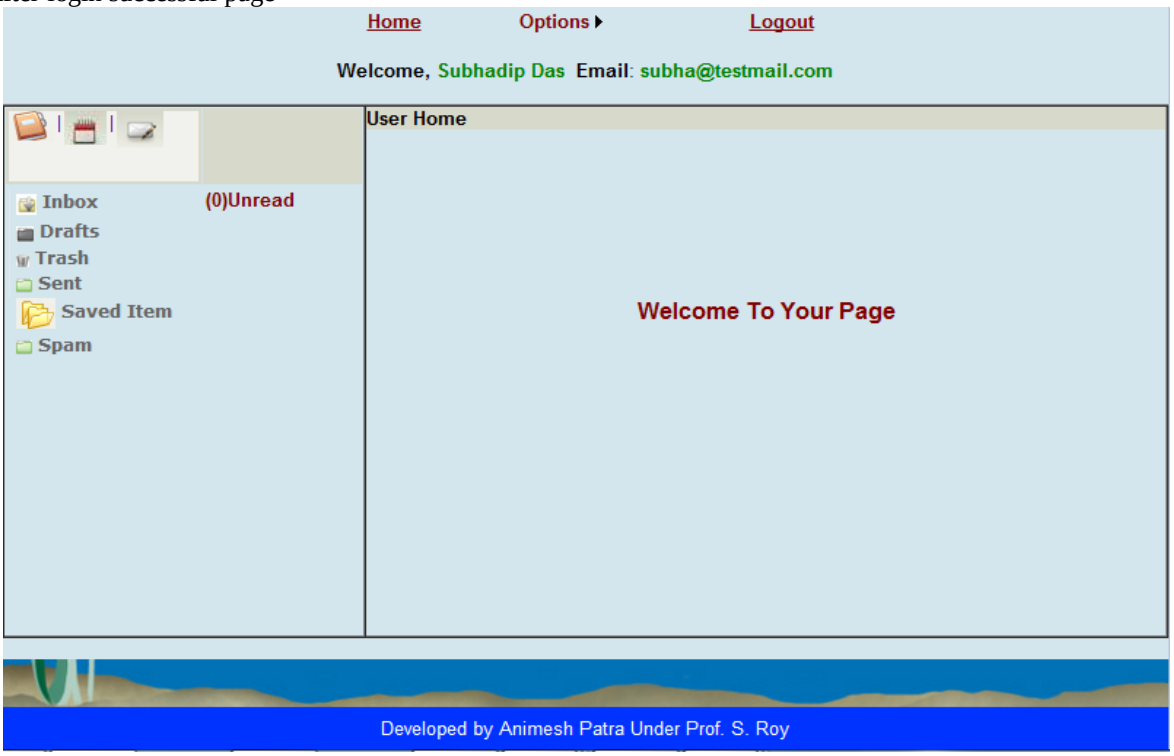

Fig. 4.5 After login to another account page

See spam folder contain the incoming email

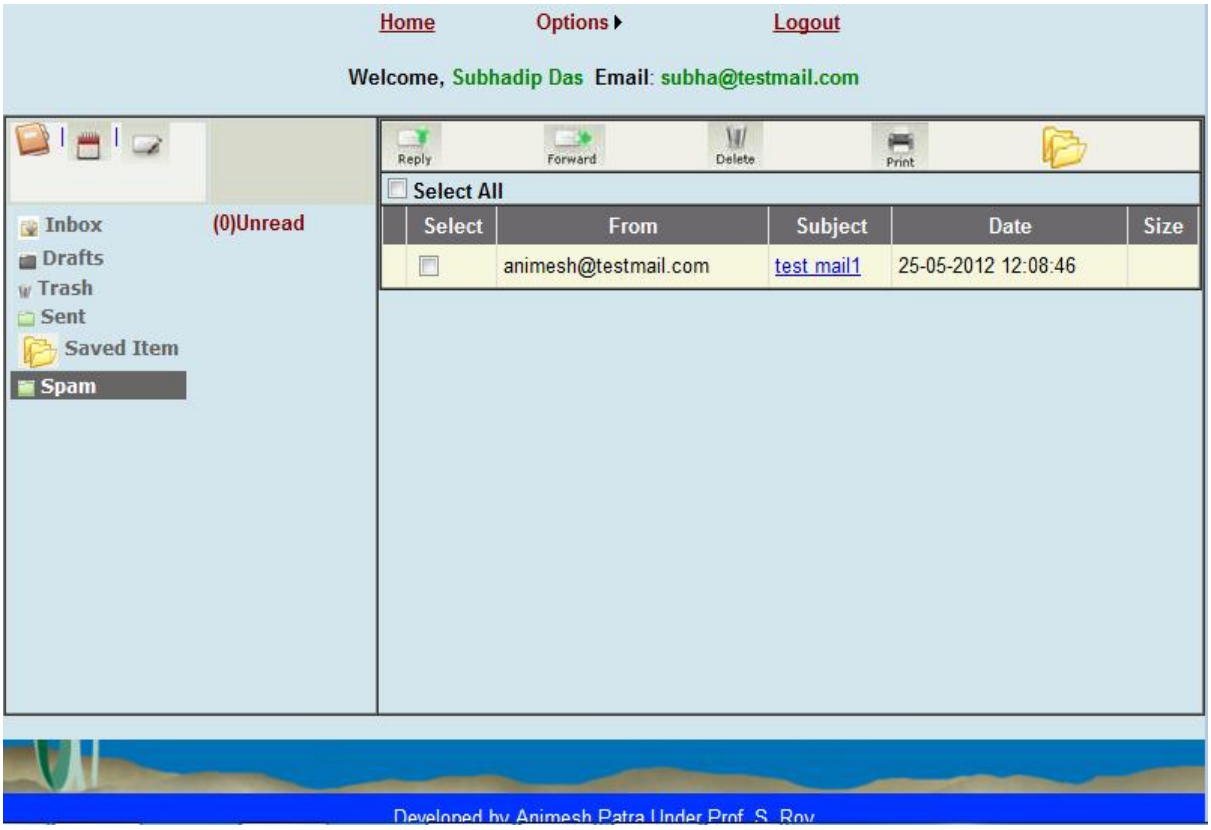

Fig. 4.6 Spam folder page

Login to another account

Fig.4.5 Another login page

Spam email content

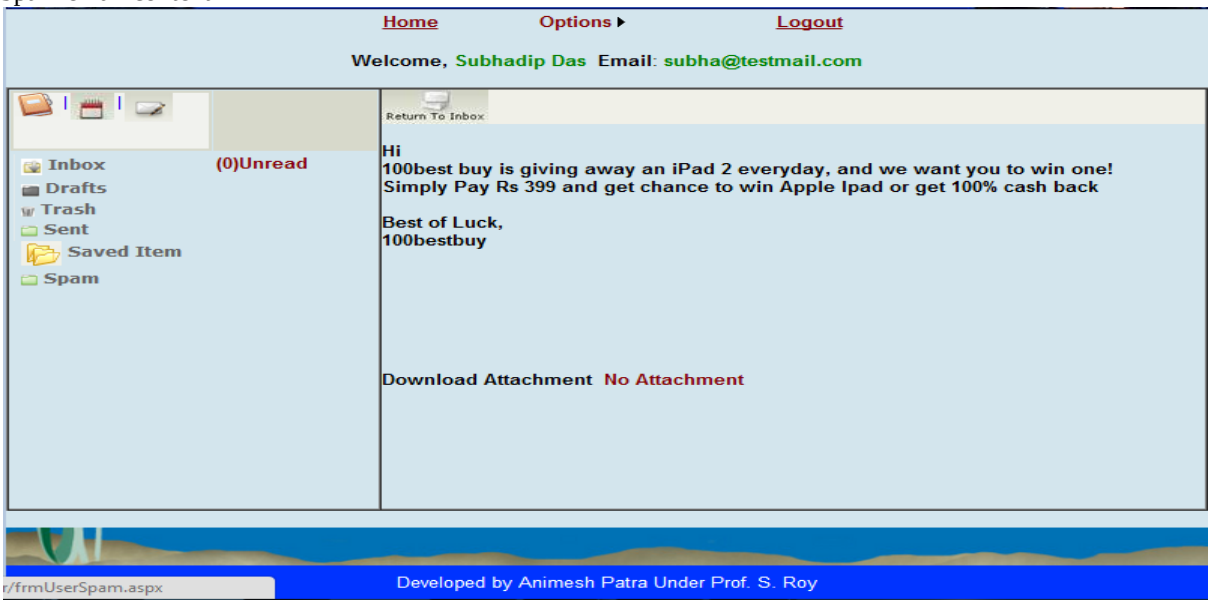

Fig. 4.7 Content of Spam Mail.

V. EXPERIMENTAL RESULT AND ANALYSIS

Estimating classifier accuracy is important in that it allows one to evaluate how accurately a given classifier will label the test data. It can be calculated using the formula discussed below. A classifier is trained to classify e-mails as non-spam and spam mails. An accuracy of 85 % may make the classifier accurate, but what if only 10-15 % 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 to as positive samples) how well it can recognize "non-spam" samples (referred to as negative samples)[8].The sensitivity (recall) and specificity measures can be used, respectively for this purpose.

The 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:

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

True Negative (TN): This states the number of non-spam documents correctly classified as non spam.

False Positive (FP): This states the number spam documents classified as non spam.

False Negative (FN): This states the number of non-spam document classified as spam.

Table.3: Measurement, Formula, and Meaning of TP, TN, FP and FN

Measurement	Formula	Meaning
Precision	TP / (TP + FP)	The percentage of positive predictions that is correct.
Recall / Sensitivity	TP / (TP + FN)	The percentage of positive labeled instances that ispredicted as positive.
Specificity	TN / (TN + FP)	The percentage of negative labeled instances that is predicted as negative.
Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions that is correct.

500 spam document and 100 non spam document are tested. From the experimental result, out of 500 spam to identify 468 as a spam and from 100 non spam document to identify 86 documents as nonspam.

True Positive (TP) = 477

True Negative (TN) = 88
False Positive (FP) = 23
False Negative (FN) = 12
Precision = TP / (TP + FP) = 0.95
Recall / Sensitivity = TP / (TP + FN) = 0.97
Specificity = TN / (TN + FP) = 0.79
Accuracy = ((TP + TN) / (TP + TN + FN + FP))*100 = 94.16%

VI. CONCLUSION

In this Dissertation, the email client system that has capability to send email and receive email and project mainly concerned about an efficient email spam filtering techniques for an email account. For this system, we collected statistical data by which we create a training set. This dataset is updated time by time. The filtering techniques based on Naive bayes Theorem, which is a good one machine learning algorithm. The projectis concentrated only on text word not any other content. But the system is very much effective to identify spam from email for text based.

REFERENCES

[1] A. Androutsopoulos, J. Koutsias, K.V. Cbandrinos and C.D. Spyropoulos. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. *In Proceedings of the 23rd ACM International Conference on Research and Developments in Information Retrieval*, Athens, Greece, 2000, 160–167.

[2] B.Klimt and Y.Yang. The Enron corpus: A new data set for e-mail classification research. *In Proceedings of the European Conference on Machine Learning*, 2004, 217–226.

[3] A. McCallum and K. Nigam. A comparison of event models for naive Bayes text classification. *In Proceedings of the AAAI Workshop on learning for text categorization*, 1998, 41–48.

[4] F.Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[5] M. Sahami, S. Dumais, D. Heckerman and E. Horvitz. A Bayesian approach to filtering junk e-mail. *AAAI Technical Report WS-98-05*, Madison, Wisconsin, 1998.

[6] L. Zhang, J. Zhu and T. Yao. An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing*, 3(4):243–269, 2004.

[7] J. Golbeck and J. Hendler. Reputation network analysis for email filtering. *In Proceedings of the First Conference on Email and Anti-spam*,2004.

[8] T.M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.