Volume: 5 Issue: XI Month of publication: November 2017
DOI:

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue XI November 2017- Available at www.ijraset.com

Image and Text Spam Mail Filtering System

Ms' Deshmukh S. S.[1], Ms. Borhade S. B[2], Ms' Swami M.M.[3], Ms.Dhole P. B. [4], Ms. Manglekar S.M[5]

Abstract: Spam mail is not only a waste of time for the recipients but can also spread harmful messages and viruses worms. It is also harmful for machine, it is not secure. In transmitting information, text is always used in image spam. Spam mail takes lot of space of inbox; it is time consuming and irritating process to manually filter spam and legitimate mail. Therefore in this study, I classify mail images by the configuration of letters text and images. In this system I have proposed a spam detection method that uses sober operators for edge detection and a multiple filter using Sobel operators and AOCR, This is beneficial for identifying spam mail. With the wide application of E-mail for communication, unwanted email means spam mail has become a major problem for E-mail users. This is also major problem for internet users.

Key terms: Spam, Ham, OCR, SVM, NB, DT, Rfn, Rfp

I. INTRODUCTION

To prevent email spa both end users and administrators of email systems use various anti spam techniques. No one technique is a complete solution to the spam problem, and each has trade-off be-tween incorrectly rejecting legitimate email vis. Common uses for mail filters include organizing incoming email and removal of spam and computer viruses. [1].The spam filtering is actually to classify the E-mails into ham and spam.

This needs to use the theory of Bayes to predict whether the received E-text mail is spam or not and it use AOCR to predict weather the received image mail is spam or not and according to the correctly classified E-mails. It consist of implementation of system which put all text and image spam emails in spam box and ham email in user inbox without manual intervention.

II. LITERATURE REVIEW

Literature review mainly divided in to two parts. Various algorithms for filtering image spam mails are NDD, SIFT, TR-FILTER, AOCR various algorithm for filtering text spam mails are as SVM,KNN,DT,BAYESIAN.[1].

A. *Various Algorithms For Filtering Text spam*

1) *Emails*

a) *SVM:* SVM is based on the structural risk minimization with the error-bound analysis. [5]. Technique of Classification: Using a kernel function, SVM are an alter-native training method for polynomial, radial basis func-tion. Less memory and time is required for SVM since in SVM we can discard all non-support vectors without any problem.

b) *KNN:* It is also known as a lazy algorithm. The ken model finds a group of k observations in the training set that are closest to the test example, and bases the assignment of the target class on the predominance of a particular class in this neighbourhood.More memory is needed as we need to store all training data. More time might be needed as in the worst case, all data points might take point in decision.[13].

c) *DT:* Decision tree is a simple structure where non-leaf nodes represent the conditional tests of attributes or features and leaf nodes contain the class label in which each data predicted into.Tree-shaped structures that rep-resent sets of decisions. These decisions generate rules for the classification of a dataset. Cost much time to build classifier. More memory is required. Stability of the tree structure is the matter of concern. [11].

d) *BAYESIAN:* Bayesian networks are graphical rep-presentation for probabilistic relationships among a set of random variables. The basic idea is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. Memory Required is Less since the complete Bayesian network structure is constructed only from the induced conditional independence and dependence information. Time Required is Less[1][2]. Particular words have particular probabilities of occurring in spam and legitimate e-mail [2],[5],[10]. The filter doesn't know these probabilities in advance, and must first be trained so that it can build them up. Bayesian filters automatically induce or learn a spam classifier from a set of manually classified examples of spam and legitimate (or ham) messages (the training collection). After training, the word probabilities (also known as liklihood functions) are used to compute the probability that an e-mail with particular set of words in it belongs to either category. Each word in the e-mail contributes to the email's spam probability, or only the most interesting words. Then, e-mail's spam probability is computed and if the total exceeds a certain threshold (say 95%), the filter will mark the e-mail as spam. The learning process takes as input the training collection, and consists of the following steps:

$$\Pr(S|W) = \frac{\Pr(W|S)\cdot\Pr(S)}{\Pr(W|S)\cdot\Pr(S)+\Pr(W|H)\cdot\Pr(H)}$$

Where
Pr(S|W) is the probability of the word.
Pr(W|S) probability that the word appears in spam mail.
Pr(S) is the overall probability that any given message is spam.
Pr(H) is the overall probability that any given message is ham.
Pr(W|H) is the probability that the word appears in ham messages.
Recent statistics show that the current probability of any message being spam is 80%, at the very least:

$$\Pr(S) = 0.8; \Pr(H) = 0.2$$

All these individual probabilities will be combined to obtain a combined probability which will be the probability of that mail. That is calculated by the formula

$$p = \frac{p_1 p_2 \cdots p_N}{p_1 p_2 \cdots p_N + (1-p_1)(1-p_2)\cdots(1-p_N)}$$

P=Probability that the message is spam.
P1=P(S|W1) (Individual probability of first word)
This combined probability of the spam mail is given to the fourth and last module i.e. Spam Detection.

B. Spam detection.

1) A threshold value for all our mails is decided.
2) Then threshold is compared with the calculated combined probability.
3) If combined probability > Threshold

Then that mail is spam.
Else that mail is ham and we put it in user’s inbox.

C. Various Algorithms for filtering Image spam Emails

1) *NDD:* Near duplicate detection Methodology Firstly, extract the features of the detected image, Secondly, compare the features of it with the features in two feature databases, by calculating their similarity, and respectively count the numbers of images that are similar to it in two DB. Finally, judge it is spam or ham by the numbers. Advantages of Near-Duplication is likely to perform well in abstracting base templates, when given enough examples of various spam templates in use .Disadvantages This technology may not be advanced enough for recognizing spam from random images with-out any explicit instructions or rules. May require user intervention. Time required Slightly more because image has to be compared with Ham and Spam dictionary
2) *SIFT:* Scale Invariant Feature Transform in this methodology When a new email comes, the filter sys-tem will extract the features of the image(s) from the e-mail and search for a matching candidate from the User-Specified Image Content(USIC) feature blacklist. If there is one, the e-mail will be blocked as spam. Other-wise it will be judged by user. The advantages of SIFT is the users are responsible for making rules for spam and ham. Disadvantages This technology may require user intervention. It also requires More time because user intervention is needed.
3) *TR FILTER:* The main idea of the text-region ex-traction method is that text often contrasts a lot with background. A region with massive intensity changes (i.e., edges) would be potentially a text region. The advantages of TR filter is its simplicity, detection using only TR-filter still achieves slightly better detection accuracy (i.e.79 percent).It Requires slightly higher computational time. Slightly lower accuracy and computational time
4) *OCR:* Optical Character Recognition (OCR) is a technology that to analyze the characters in pictures and convert them to texts. OCR is software which takes image as an input and recognizes the text in the given image as an output. For character recognition, offline or online, there are two basic types of core OCR algorithm. Optical Character Recognition (OCR) is a

process of converting printed or handwritten scanned documents into ASCII characters that a computer can recognize. In other words, automatic text recognition using OCR is the process of converting an image of textual documents into its digital textual equivalent. Developing an OCR is a very difficult task, It can be used in detecting spam mail by checking the worlds in the image.

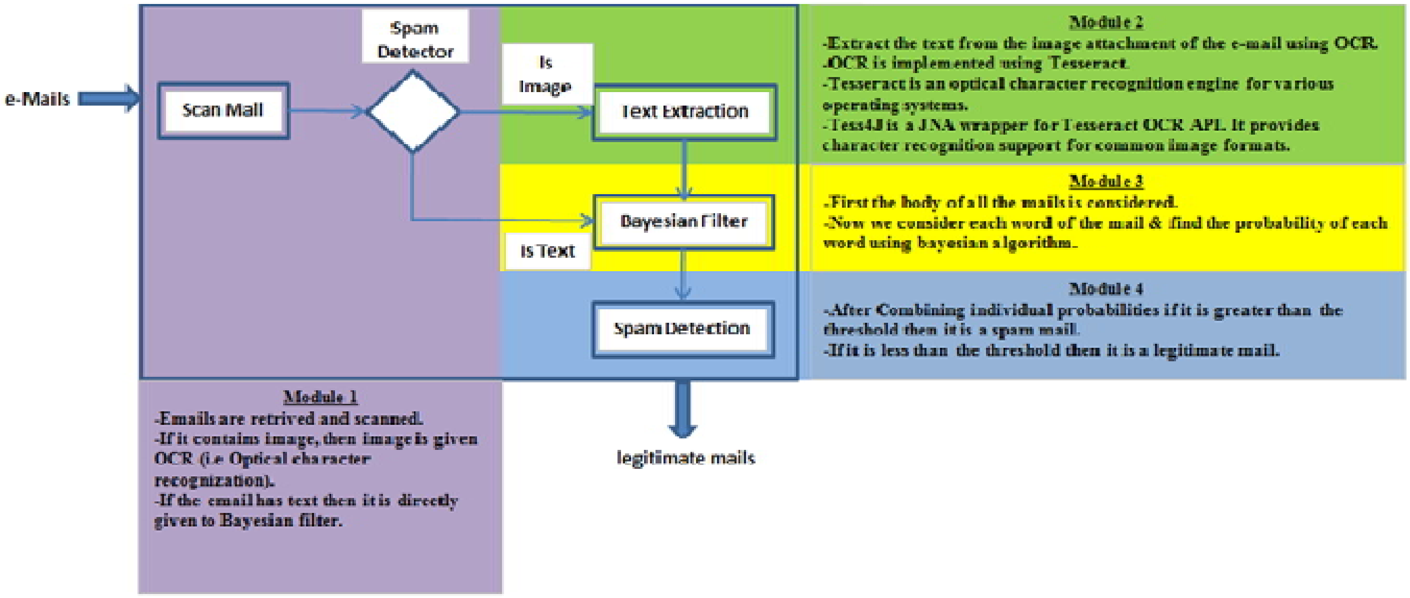
D. *Advantages of Bayesian Algorithm*

Advantage of Bayesian classification Compared to other classification filtering, Bayesian classification method has the following advantages

1) Better than the other algorithms in efficiency. Bayesian classification algorithm to scan all the training samples again, and statistics for each word in the normal email and spam in the number of occurrences of each Token after the query just once again, the last Token for each product or additive. The SVM method requires scanning multiple training samples.
2) In storage, Bayesian classification algorithm only needs to store the number of words, rather than the actual message. Thus, very little storage space, but the resulting data can be shared between users without considering privacy of message.
3) Bayesian classification methods continue to receive a single message with the incremental update, you can adapt to the evolution of forms of spam. Changes in the content of spam was more, Bayesian classification methods can be collected from users under the guidance of recently received spam features, effectively

III. SYSTEM ARCHITECTURE

The proposed system takes input as a user mails. User mail is input to the text or image identification block. In this block a decision is taken weather the mail contains text or image. If the mail contains text then it is directly given to Bayesian filter else the text from image is detected and then it is given to spam detection block.



Module 2
-Extract the text from the image attachment of the e-mail using OCR.
-OCR is implemented using Tesseract.
-Tesseract is an optical character recognition engine for various operating systems.
-Tess4J is a JNA wrapper for Tesseract OCR API. It provides character recognition support for common image formats.

Module 3
-First the body of all the mails is considered.
-Now we consider each word of the mail & find the probability of each word using bayesian algorithm.

Module 4
-After Combining individual probabilities if it is greater than the threshold then it is a spam mail.
-If it is less than the threshold then it is a legitimate mail.

Module 1
-Emails are retrived and scanned.
-If it contains image, then image is given OCR (i.e Optical character recognization).
-If the email has text then it is directly given to Bayesian filter.

1) This is the system architecture diagram of our system.

It has total 4 modules scan mail, text extraction, Bayesian filter, spam detection.

Input-All mails

Output -Legitimate mails.

*2) Module 1:*Scan mail

a) This module will accept the username and password from the user and login to our system.

b) Then the system is interfaced to the server (ex. Gmail) and mails are retrieved.

c) If the mail contains image then it is given to model 2 i.e. Text extraction and if it contains only text it is given to model 3 i.e. Bayesian filter.

3) Module 2: Text extraction

a) Text is extracted from image in this module with the help of OCR (optical character recognition) which is a ready software. OCR is implemented using tesseract.

b) The extracted text is then given to the module3 i.e. Bayesian filter.

4) module 3: Bayesian filter

a) This is the main module of our system.

b) In this algorithm we consider only the body of our mail.

c) We find out the probability of each word in the body of an email by the formula

pspam=rbad/rbad+rgood

Where,

pspam is the probability of the word.

rbad is the probability of that word in spam database

rgood is the probability of that word in ham database

4) *All these individual probabilities will be combined to obtain a combined probability which will be the probability of that mail. That is calculated by the formula*

pspam=pposproduct / pposproduct + pnegproduct

Where,

pspam is the combined probability of the mail

pposproduct is the product of all the individual probabilities of the words in the mail

i.e. pposproduct=pspam1 * psapm2 *....*pspamn

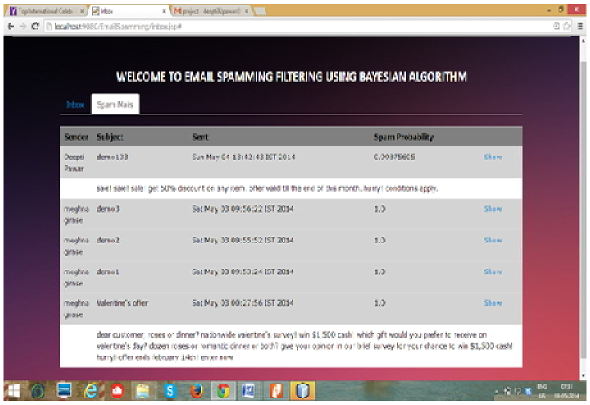
where, psapm1 is the probability of 1st word in the mail and so on.

pnegproduct= (1-pspam1) * (1-pspam2)*...*(1-pspamn)

5) *Module 4:*Spam detection.We already consider a threshold value for all our mailsWe compare the calculated combined probability with the threshold valueIf combined probability > ThresholWe say that mail is spam.Else we conclude that mail is ham and we put it in user's inbox.

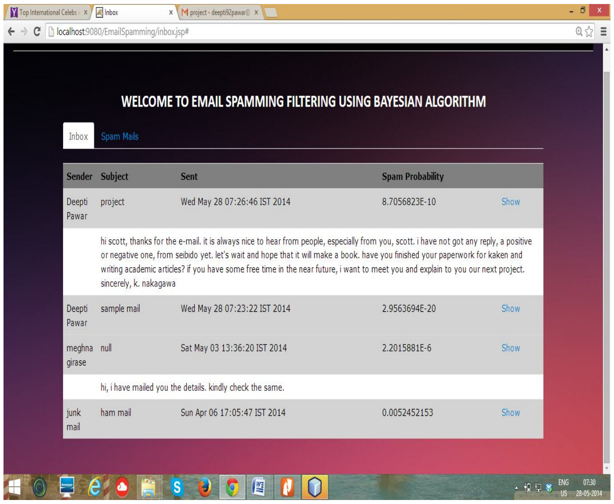
IV. EXPECTED RESULT DISCUSSION

Image Text Spam mail filtering system takes user mail as input, the mail can have text as well as image. In this system the text from image is identified with the help of AOCR Bayesian is useful to detect the words in email are spam or not. Expected result is email is classified as spam or ham without user interaction. This system currently works on one single machine, in future it will work on network.



WELCOME TO EMAIL SPAMMING FILTERING USING BAYESIAN ALGORITHM

Sender	Subject	Sent	Spam Probability	
Deepti Pawar	demo133	Sun May 04 13:42:43 IST 2014	0.99379560S	Show
	sale! sale! sale! get 50% discount on any item. offer valid till the end of this month. hurry! conditions apply.			
meghna girase	demo3	Sat May 03 09:56:22 IST 2014	1.0	Show
meghna girase	demo2	Sat May 03 09:55:52 IST 2014	1.0	Show
meghna girase	demo1	Sat May 03 09:53:24 IST 2014	1.0	Show
meghna girase	Valentine's offer	Sat May 03 09:27:56 IST 2014	1.0	Show
	dear customer, roses or dinner? nationwide valentine's survey! win $1,500 cash! which gift would you prefer to receive on valentine's day? dozen roses or romantic dinner or both? give your opinion in our brief survey for your chance to win $1,500 cash! hurry! offer ends february 14th! enter now			



WELCOME TO EMAIL SPAMMING FILTERING USING BAYESIAN ALGORITHM

Sender	Subject	Sent	Spam Probability	
Deepti Pawar	project	Wed May 28 07:26:46 IST 2014	8.7056823E-10	Show
	hi scott, thanks for the e-mail. it is always nice to hear from people, especially from you, scott. i have not got any reply, a positive or negative one, from seibido yet. let's wait and hope that it will make a book. have you finished your paperwork for kaken and writing academic articles? if you have some free time in the near future, i want to meet you and explain to you our next project. sincerely, k. nakagawa			
Deepti Pawar	sample mail	Wed May 28 07:23:22 IST 2014	2.9563694E-20	Show
meghna girase	null	Sat May 03 13:36:20 IST 2014	2.2015881E-6	Show
	hi, i have mailed you the details. kindly check the same.			
junk mail	ham mail	Sun Apr 06 17:05:47 IST 2014	0.0052452153	Show

V. CONCLUSION

This system is developed to text and image spam emails. It extracts the body of the email and using Bayesian algorithm it finds out whether the email is spam or legitimate mail. Unlike the current technique (DNSBLs, used to publish the addresses of computers or networks linked to spamming; most mail server software can be configured to reject or flag messages which have been sent from a site listed on one or more such lists), in image and text spam mail filtering system the focus is on the content of the email. If the email contains spam words then only it is labeled as spam.

This system works only for text emails and the emails that contains image, pdf, text file as an attachment. It is not applicable to any zipped file. It only focuses on the content of the email and not on the subject or URLs present in the emails.

REFERENCES

[1] Hu Yin, Zhang Chaoyang,An Improved Baysian Agorithm for filtering spam E-mail,2011 Journal of Computational Information Systems (2011)

[2] Yishan Gong, Qiang Chen,Research of Spam Filter-ing Based on Bayesian Algorithm2010 International Conference on Computer Application and System Modeling.

[3] Christina V, KarPagavalli S, SUganya G.A Study on Email Spam Filtering Techniques. International Jour-nal of Computer Applications (0975 - 8887),Volume 120 - No.1. December 2010

[4] Mori Tatsuya: On the use and misuse of E-mail sender authentication mechanisms, IEICE technical report 110(115), pp.101-106, 2010.

[5] Mori Tatsuya: PrBL: Probabilistic Blacklist for E-mail Spammers, IEICE technical report 108(457), pp.15-20, 2009.

[6] Wang Meizhen, Li Zhitang, Wu Hantao.An im-proved Bayes algorithm for filtering spam e-mail. J. Huazhong Univ. of Sci. Tech(Natural Science Edi-tion), Vol 37 No 8. Aug 2009

[7] Ravi Kiran S S, Indriyati Atmosukarto. Spam or Not Spam-That is the question. 2009

[8] Liu Pei-yu,Zhang Li-wei,Zhu Zhen-fang,Research of Email Filtering Based on Bayesian,Journal of Com-puter, vol. 4,no. 3,march 2009.

[9] Johan Hovold, Naive Bayes Spam Filtering Us-ing World-Position-Based Attributes, Department of Computer Science,Lund University 2008.

[10] White Paper- Why Bayesian filtering is the most effective anti-spam technology.2008

[11] Sugii Manabu, Matsuno Hiroshi: Decision Tree Representation of Spam Mail Features by Machine Learning, IPSJ SIG Notes 2007(16), pp.183-188, 2007.

[12] Guy Di Mattina: Spam and Open Relay Blocking System, A thesis submitted to the School of Infor-mation Technology and Electrical Engineering The University of Queensland, 2003.

[13] C. Romero, M. Garcia Valdez, A. Alanis,A Compar-ative Study of Machine Learning Techniques in Blog Comments Spam Filter