Global Journal of Advanced Engineering Technologies *Volume 5, Issue 2- 2016*
ISSN (Online): 2277-6370 & ISSN (Print):2394-0921

IMAGE AND TEXT SPAM MAIL FILTERING

[1]*Nisha D. Chopra* , [2] *Prof. I.R. Shaikh*
[1] *Student, Department of Computer Engineering, SND COE and RC Yeola, Nashik, India.*
[2]*Associate Professor, Department of Computer Engineering, SND COE and RC Yeola, Nashik, India.*

Abstract: *E-Mails the fastest and the easiest way of E-Communication, brings with it some threats to the system of Internetwork. There are many unethical personalities who are continuously trying to hamper the smooth working of internet. This is done by continuously sending a bulk of messages, which are meaningless or may contain some virus files. With this the both the internet and the receiver system goes down. For this, there needs a system which detects such unimportant messages and performs appropriate action on it.*
Simply explaining, the bulk of messages send by the sender are the SPAM mails, and the sender sending the spam mails is called the SPAMMER. Now there is a requirement of developing the system which is able to detect the spam mails coming to the receiver and stops the spammers to it further. This means, the need to develop an anti-spam mail filter, is the obvious need.
Current spam mail filters are developed such that they can detect various properties of spam mail. Specially, text categorization technique is used for filtering the spam mails. However, spammers introduced the new method of dumping the spam mails in the attached image in the mail. In this paper we, put forward a method of departing the text from the image, and then, detecting the spam text message from the mail list.
Keywords: *spam filtering, e-mail, images, text categorization, Optical Character Recognition.*

I. INTRODUCTION

Since last many years, the continuous growth of spam mails that is, the bulk delivery of unimportant emails, mainly of commercial nature or with non needed content. This can become a main problem of the email service for Internet service Provider.
The recent study survey of email server, has reported 60% of all email traffic is the Spam [1].
The image spam mails traffic is SPAM too. The image spam mails have added a number to this list. The spam mails cause, two severe problems, first one is, it overloads the bandwidth and second is, it also burdens the storage capacity of the server. This results in unnecessary time and storage memory consumption and also raises the annual cost of the communication.
The spam mail is in addition a serious threat for the security of the end email service users, since the spammers try to learn and occupy important personal information like account no. and passwords.
As in so short period it is impossible to change the internet protocols for such mails categorization, the only solution is developing different way for filtering spam mails.

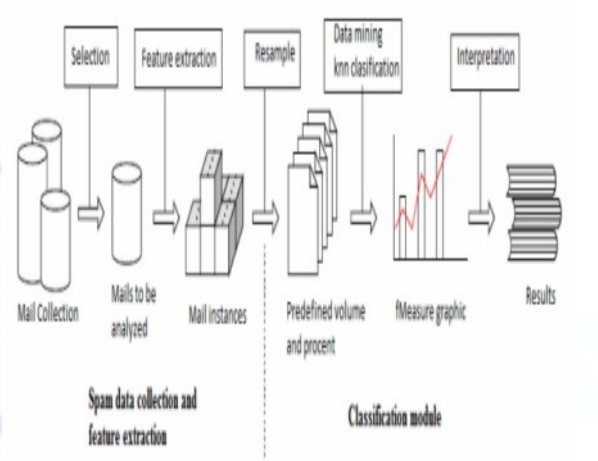

Figure 1: Architecture of the system

Here, the paper raises forward an image spam detection system to fulfil the requirements. The basic idea is to use traditional anti-spam methods to detect some spam messages that are image based. The system we propose is based on two remarks [2]. The first is that traditional spam detection methods such as honey- pots, message header analysis or human reporting procedure can detect some spam messages image based. The second is that image spam messages are basically sent in large heaps where the messages in each heap are visually similar, although the variations can be sophisticated. For example, spammers frequently design a template image and apply various alterations or noise to the template before sending it to each of the end user in target. It is believed that this is because spammers still want the clear information to be delivered to the end user and they want to use ancient methods to produce

Global Journal of Advanced Engineering Technologies
ISSN (Online): 2277-6370 & ISSN (Print):2394-0921
Volume 5, Issue 2- 2016

millions of exclusive spam images while not puzzling the template image too much.

II. LITERARTURE SURVEY

Under the present situation of internet, plenty of spam mail filters are widely set up, for example Mail Avenger, Spam Guru Etc. These anti-spam techniques start by tracking the source, destination, path and various type of information for spam analysis. Many of these spam mail filtering systems segregate these messages into White List, Black List and Grey List based on the contents of the messages. Some anti-spam systems also use the technique of block list which blocks the spam mails sender. Many of the systems such as Mail Avenger, Spam Assassin use many of the techniques for filtering text based spam, including plain text and HTML elements. Traditional spam filters, totally depend on examining of sender, message header, path etc. other information can detect some of the image spam without visualizing the whole image itself.

A variety of technical processes against spam email have been already projected: decision trees (DT), support vector machines (SVM), K-nearest neighbour algorithm, naive Bayes (NB), neural networks, ensemble decision trees (EDT), boosting, bagging and stacking etc. Most of the techniques above can be effectively applied to the problem of spam mails, but among all these algorithms, content based filtering (and in particular, Bayesian filtering) is playing an important role in eliminating spam email. The spam filtering is actually used to sort the E-mails into HAM and SPAM. This needs to use the theory of Bayesian to predict whether the received E-mail is spam or not, according to the correctly sorted E-mails.

III. DETAILS OF DISSERTATION WORK

The system projected here, in this paper, is able to identify and filter both, the text spam mails and also the text implanted in the image. To do this for text based spam mails, the Bayesians filter Algorithm is used and for the image based spam mail filters the effective OCR algorithm is used.

The OCR filters are highly valuable for clean images. This proposed system uses classification of images methods to classify between Ham and Spam images, which use low level visual characteristics related for an instance to colour distribution, characteristics of text regions inside an image. With this, there will be un effect of text obfuscation techniques used by spammers. The overall computational cost of the proposed system is lower than other simple OCR-based approaches.

The proposed system also gives the provision to the end user to delete or block the sender of the spam mails.

A) Mathematical Model

The system projected in this paper is the anti spam system. As soon as the new mail is received by the user in his Gmail account, the system reads the mails and if it contains any text embedded in image, it works by first extracting the text from the image. The extracted text is then examined by the Bayesians algorithm for Spam or Ham.

- *Objective: To filter mails and to provide legitimate mails to user.*
- *Actors: Hammer, Spammer, User, and System.*
- *Input: Emails.*
- *Output: Legitimate mails, Spam mails.*

Problem Statement:

Let S be the mail filtering system such that,
S= {M, U, F, OT}
M represents the set of mails. M = {L, Sm}
L represents the Set of Legitimate mails. L = {L1, L2… Ln}
Sm represents the Set of Spam mails. Sm = {Sm1, Sm2, ….. Smm}
U represents the set of User. U = {U1, U2… Up}
F represents the Filter. F = {B, O}
Where,
B = Bayesian Filter, O = OCR filter.
OT represents the output i.e. the set of legitimate mails.
OT = {OT1, OT2… OTr}

B) Data Flow Diagram

Level 0 DFD with the figure below shows the flow of user accessing the Gmail account by giving the user name and password of the Gmail account. The system provides with the legitimate mails in the inbox of the user.

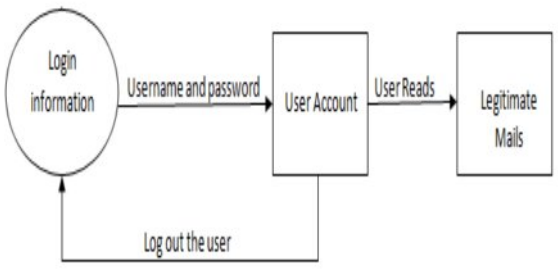

Figure 2: Level 0 data flow diagram

Global Journal of Advanced Engineering Technologies
ISSN (Online): 2277-6370 & ISSN (Print):2394-0921
Volume 5, Issue 2- 2016

i) Level 1 Data Flow Diagram

Level 1 DFD describes the flow of the system, as the user provides the system with his correct authentication information; he gets access to the system, where the system reads the incoming mail. The system detects and recognizes the mail is spam or not.

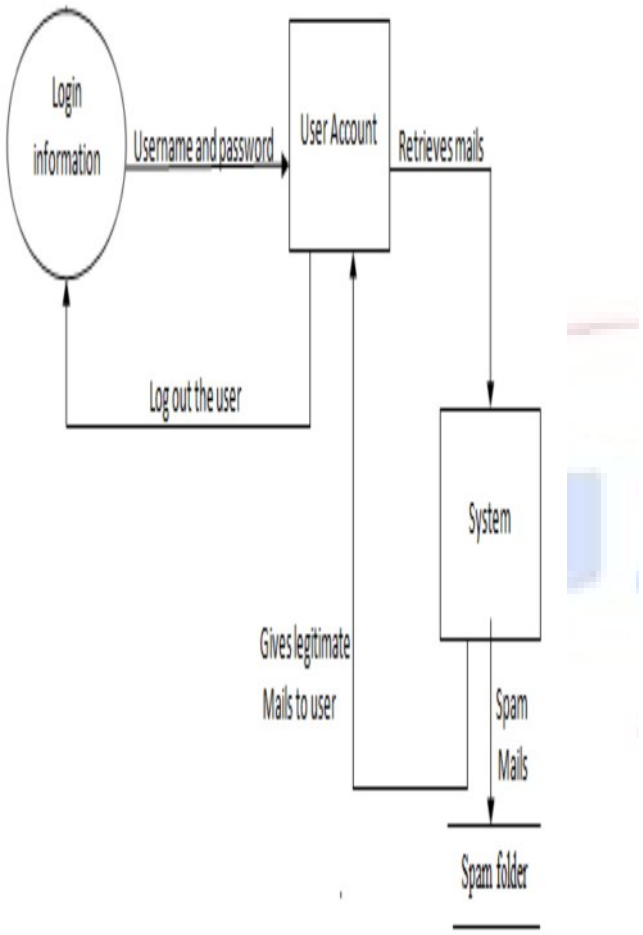

Figure 3: level 1 Data flow diagram

ii) Level 2 Data Flow Diagram

Level 2 DFD show the detailed working of the proposed system. The system here inspects the inbox new received mail and if the content is image, it is then given to OCR Algorithm. If the content is text then it is given to Bayes algorithm.

Level 2 data flow diagram

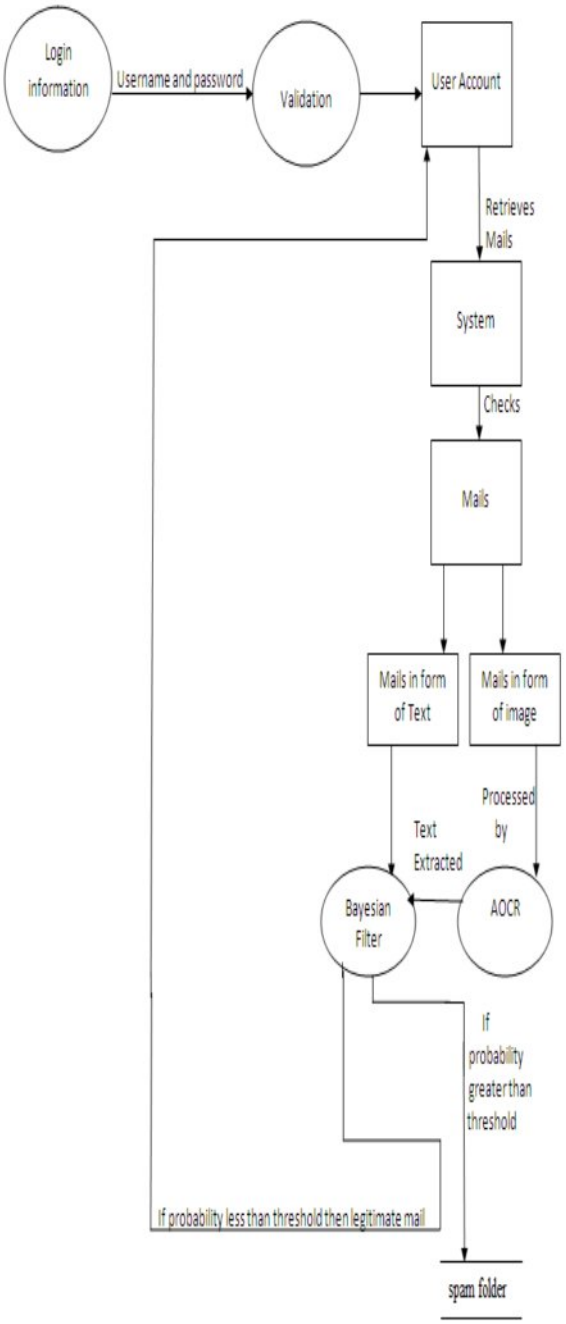

Figure 4: Level 2 DFD

Global Journal of Advanced Engineering Technologies Volume 5, Issue 2- 2016
ISSN (Online): 2277-6370 & ISSN (Print):2394-0921

C) System Architecture:

The system architecture for the proposed system is shown below [5].The architecture is separated into 4 fundamental modules depicted in four different colours in the figure.

Module 1:

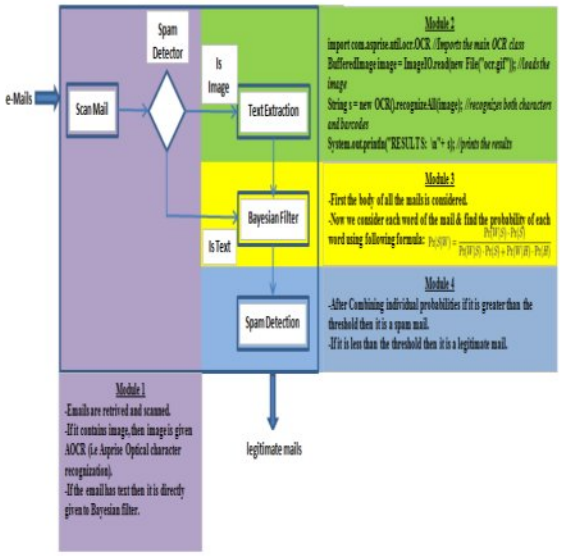

Fig 5: System Architecture.

The purple block of the system architecture shown above is the first module. This module reads the new inbox mails. It also reads the mails and sends it to the detector of the system, which detects whether the input mail contains the image or just the text.

The Second module is shown with green colour. This module takes the input from the spam detector of the first module. This module basically performs the text extraction from the image of the input mail.

The Bayesians formula is used in this proposed system for text mails classification into spam or ham. The third module consists of a Bayesian filter. The input the third module is from two sources, ie the first or the second module. This filter uses the Bayesian formula to calculate the total probability of the spam words in the mails.

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

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.*

The Fourth Module is the last block of the system. This module actually shows the output result of the system. From the total calculated probability of the mail, comparison is done with the pre decided threshold value. As the calculated probability exceeds the threshold value, the output of the system is the SPAM MAIL or else the output is the HAM MAIL which is stored in the inbox.

IV. RESULTS AND DISCUSSION

A) Result set

Features	*Algorithm*	*Results*
No. Of users/ Response Time	*OCR & Bayesians* *OCR & SVM*	*25 secs per mail* *35 secs per operation.*
Accuracy Factor	*OCR & Bayesians* *TR Filter* *OCR & SVM*	*8 out of 10* *6 out of 10* *9 out of 10*

The system has been tested on three systems simultaneously; the result obtained is depicted in the above table. The projected system uses the OCR and the Bayesians algorithm.

As shown in the table, the proposed system takes just 25 secs of time to detect the new inbox mails, and classifies it as spam or legitimate mail.

When proposed system was executed for 10 mails including spam and legitimate mails, the system was able to detect 8 correct classifications of input mails.

V. CONCLUSION

From this system one can get rid of spam mails from the spammer, thus reducing the memory and time utilization. System is providing a special spam folder which stores all the spam detected mails in that folder. System also provides many constraints to be put on the spam mail sender with the user intervention. The result provided by the system is a table providing a detail of all the inbox mails, filtration done on it and the action performed accordingly.

VI. FUTURE SCOPE

Thus, the system created can easily be extended to other email servers and can be made platform independent. With this approach one can easily and effectively protect the confidential and important email account from the spammers. The system can further be enhanced by embedding many actions on the email sender or the overall inbox. With this,
many privileges and constraints can be applied to the sender of the email depending the mail is spam mail or legitimate ham mail.

REFERENCES

[1] Ying Tan, Guyue Mi, Yuanchun Zhu, and Chao Deng, Artificial Immune System Based Methods for Spam Filtering Key Laboratory of Machine Perception (Ministry of Education) Department of Machine Intelligence, School of Electronics Engineering and Computer Science Peking University, China.

[2] B. Fadiora, F. Wada O.B. Longe, "Combining Optical Character Recognition(OCR) and Edge Detection Techniques to Filter Image-Based Spam" Department of Computer Science The Polytechnic Ibadan Ibadan, Nigeria.

[3] Mark Dredze,Reuven Gevaryahu,Ari Elias-Bachrach, "Learning Fast Classifiers for Image Spam" Computer and Information Sciences Dept.University of Pennsylvania Philadelphia.

[4] Zhe Wang, William Josephson, Qin Lv, Moses Charikar, Kai Li, "Filtering Image Spam with Near-Duplicate Detection" Computer Science Department, Princeton University 35 Olden Street, Princeton.

[5] V. Sathiya, M.Divakar, T.S. Sumi, "PARTIAL IMAGE SPAM E-MAIL DETECTION USING OCR" International Journal of Engineering Trends and Technology- May to June Issue 2011,Chennai, India.

[6] Priyanka Sao, Pro. Kare Prashanthi, E-mail Spam Classification Using Nave Bayesian Classifier, International Journal of Advanced Research in Computer Engineering Technology (IJARCET) Volume 4 Issue 6, June 2015.

[7] Deshmukh Swati S, Prof. Chandre P.R , "Survey on: Naive Bayesian and AOCR Based Image and Text Spam Mail Filtering System",International Journal of Emerging Technology and Advanced Engineering (Certified Journal, Vol 4, Issue 4, April 14) Text and Image Spam Mail Filtering

[8] Hu Yin , Zhang Chaoyang, An improved Bayesian Algorithm for Filtering Spam E-mail, 2011 International Symposium on Intelligence Information Processing and Trusted Computing.

bm Giorgio Fumera FUMER,Ignazio Pillai PILLAI,Fabio Roli, "Spam Filtering Based On The Analysis Of Text Information Embedded Into Images" , Dept. of Electrical and Electronic Eng. University of Cagliari Piazza dArmi, 09123 Cagliari, Italy.

[9] Yishan Gong, Qiang Chen, Research of Spam Filtering Based on Bayesian Algorithm, 2010 International Conference on Computer Application and System Modeling.

[10] Christina V, KarPagavalli S, SUganya ," A Study on Email Spam Filtering Techniques", International Journal of Computer Applications, December 2010.

[11] Wang Meizhen, Li Zhitang, Wu Hantao, "An improved Bayes algorithm for filtering spam e-mail", J. Huazhong Univ. of Sci.

[12] Jianyi Wang Kazuki Katagishi, Image Content-Based Email Spam Image Filtering, Journal of Advances in Comp Networks, Vol. 2, June 14.

[13] Aris Kosmopoulos,Georgios Paliouras, Ion Androutsopoulos, "Adaptive Spam Filtering Using Only Naive Bayes Text Classifiers",Athens University of Economics and Business, Athens, Greece.

[14] Zhaoyang Qu, Yingjin Zhang, Filtering Image Spam using Image Semantics and Near-Duplicate Detection, 2009 Second International Conference on Intelligent Computation Technology and Automation.

[15] Meghali Das and Vijay Prasad, "ANALYSIS OF AN IMAGE SPAM IN EMAIL BASED ON CONTENT ANALYSIS", International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June 2014.

[16] Christina V, Karpagavalli S, Suganya G, " A Study on Email Spam Filtering Techniques", International Journal of Computer Applications (0975 8887) Volume 12 No.1, December 2010

[17] Mallikka Rajalingam, Putra Sumari, Valliappan Raman, "Text Detection and Extraction from Document Images using K-Nearest Neighbor Rule" , International Journal of Computer and Information Technology Volume 03 Issue 04, July 2014.

AUTHORS

N. D. Chopra, Post Graduate Student, was with Pune University, Maharashtra, India. She is now with the Department of Computer Engineering, SND COE, Pune University, Maharashtra, India.
(E-mail:janhavi.rawal24@gmail.com).

Prof. I. R. Shaikh, is with the Computer Engineering Department, University of Pune, Maharashtra, India.