Spam Email Classification & Blocking

Prof. K. S. Warke[1] NilamSalunkhe[2] PallaviKamble[3] Shweta Kulkarni[4] Pooja Nage[5]
[2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]BV College of engineering for women, Pune, India

Abstract— Spam emails are the emails receiver does not wish to receive; it is also called unsolicited bulk email. Emails are used daily by number of user to communicate around the world. Today large volumes of spam emails are causing serious problem for Internet user and Internet service. Such as it degrades user search experience, it assists propagation of virus in network, it increases load on network traffic. It also wastes user time, and energy for legitimate emails among the spam. It is time consuming and laborious to remove spam email by hand if there are too many spam.

Key words: Spam Email, network traffic

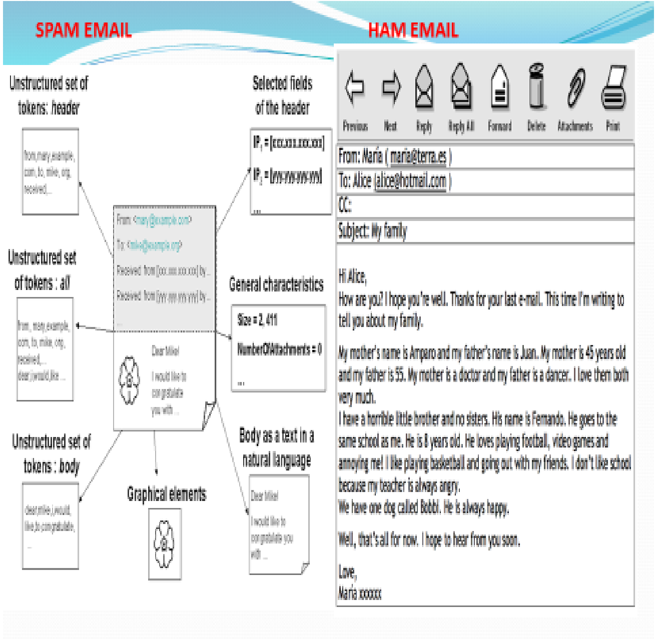
I. INTRODUCTION

A. *Data Mining:*

Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behaviour of their customers and potential customers. It discovers information within the data that queries and reports can't effectively reveal. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information thatcan be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users toanalyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

B. *Spam E-Mail:*

The problem of undesired electronic messages is nowadays a serious issue, as spam constitutes up to 75{80% of total amount of email messages. Spam causes several problems,some of them resulting in direct financial losses. More precisely, spam causes misuse of traffic, storage space and computational power. spam makes users look through and sort out additional email, not only wasting their time and causing loss of work productivity, but also irritating them and, as many claim, violating their privacyrights .



From: María (maria@terra.es)
To: Alice (alice@hotmail.com)
CC:
Subject: My family

Hi Alice,
How are you? I hope you're well. Thanks for your last e-mail. This time I'm writing to tell you about my family.

My mother's name is Amparo and my father's name is Juan. My mother is 45 years old and my father is 55. My mother is a doctor and my father is a dancer. I love them both very much.
I have a horrible little brother and no sisters. His name is Fernando. He goes to the same school as me. He is 8 years old. He loves playing football, video games and annoying me! I like playing basketball and going out with my friends. I don't like school because my teacher is always angry.
We have one dog called Bobbi. He is always happy.

Well, that's all for now. I hope to hear from you soon.

Love,
María xxxxx

Fig. 1: Structure of Spam email &Ham email

II. THE PROPOSED APPROACH

Following figure shows the system architecture of our project.

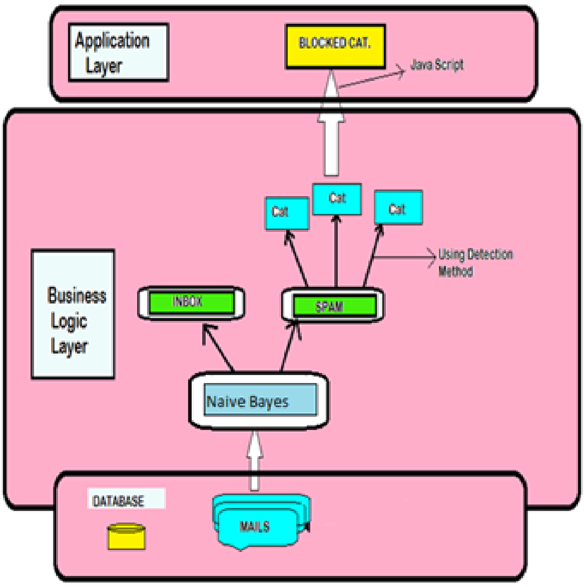


Fig. 2:

In this proposed approach first we create a database of good e-mails and spam emails in mysql. Then apply Naïve Baye's algorithm for classify the emails in spam and non spamcategory. Spam e-mails are classified into various categories by using Bayesian detection method. After that we are block and unblock the specific types of spam email.

A. Naïve Bayes Algorithm:

1) Training /Learning:

- For each mail, specify its category manually.
- Divide the mail into tokens (both subject and body)
- Filter out stop words such as html tags, articles, proverbs, noise words.
- Extract keywords and store them along with frequency count in to keywords database of the selected category.

2) Classification:

- For each newly arrived mail, divide the mail into set of tokens. (Consider both, the subject and the body)
- 2. Filter out stop words such as html tags, articles, proverbs, noise words and extract the keywords, say E {e1,e2, e3…en} is the list of extracted key words.
- Find P(category| E) = P(E | category)*P(category)/ P(E) for all categories where , P (E | category) = P(e1|category)*P(e2|category)*…..*P(en | category)
- Find the category for which the value of P(category | E) is highest.
- Compare the value with threshold value of that category. If P(category | E) > threshold , the mail is classified into that category.

B. Spam Filtering:

In this paper,we are using The Naïve Bayes Classifier method for filtering Spam emails.

The Naïve Bayes Classifier method:

The classical naïve Bayesian approach was used to develop the spam filter.

The use of naïve Bayesian classifier has become highly prevalent as the ensuing system will be less complex. Naïve Bayesian classifier is a probabilistic classifier based on Bayes' theorem. The theorem assumes that each feature is conditionally independent of each other. In 1998 the Naïve Bayes classifier was proposed for spam recognition.

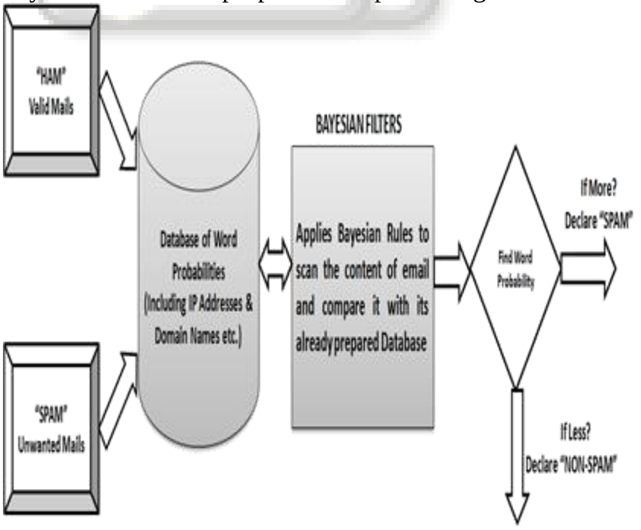

"HAM"
Valid Mails
"SPAM"
Unwanted Mails
Database of Word Probabilities (Including IP Addresses & Domain Names etc.)
BAYESIAN FILTERS
Applies Bayesian Rules to scan the content of email and compare it with its already prepared Database
Find Word Probability
If More?
Declare "SPAM"
If Less?
Declare "NON-SPAM"

Fig. 3: Bayesian Filter Method

Bayesian classifier is working on the dependent events and the probability of an event occurring in the future that can be detected from the previous occurring of the same event.This technique can be used to classify spam e-mails; words probabilities play the main rule here. If some words occur often in spam but not in ham, then this incoming e-mail is probably spam. Naïve bayes classifier technique has become a very popular method in mail filtering software. Bayesian filter should be trained to work effectively. Every word has certain probability of occurring in spam or ham email in its database. If the total of words probabilities exceeds a certain limit, the filter will mark the e-mail to either category. Here, only two categories are necessary: spam or ham. Almost all the statistic-based spam filters use Bayesian probability calculation to combine individual token's statistics to an overall score and make filtering decision based on the score. The statistic we are mostly interested for a token T is its spamminess calculated as follows :

$$\mathbf{S\ [T]} = \mathbf{C}\ \mathrm{Spam}\mathbf{(T)\ /C}\ \mathrm{Spam}\mathbf{(T)} + \mathbf{C}\ \mathrm{Ham}\mathbf{(T)}$$

Where CSpam(T)and CHam(T)are the number of spam or ham messages containing token T, respectively. To calculate the possibility for a message M with tokens {T1,......,TN}, one needs to combine the individual token's spamminess to evaluate the overall message spamminess. A simple way to make classifications is to calculate the product of individual token's spamminess and compare it with the product of individual token's hamminess.

```
Stage1. Training
Parse each email into its constituent tokens
Generate a probability for each token W
S[W] =  Cspam(W) / (Cham(W) + Cspam(W))
store spamminess values to a database.
Stage2. Filtering
For each message M
while (M not end) do
scan message for the next token Ti
query the database for spamminess S(Ti)
calculate accumulated message probabilities
S[M] and H[M]
Calculate the overall message filtering indication by:
I[M] = f(S[M] , H[M])
f is a filter dependent function,
such as,
I [M] = 1+S[M]-H[M]/2
if I[M] > threshold
msg is marked as spam
else
msg is marked as non-spam
```

III. BLOCKING &UNBLOCKING

In this, we are using some standard and popular algorithm Naïve bayes for spam email classification. The best worldwide Companies which provides email services like Google, Yahoo, Rediffmail etc. are not providing Blocking Facility of Spam email. So there is a Need of Spam email blocking in our daily routine.

We are using the Java script for blocking spam emails. But some of the spam emails are important so we are providing Unblocking Facility also.

IV. Result of Project

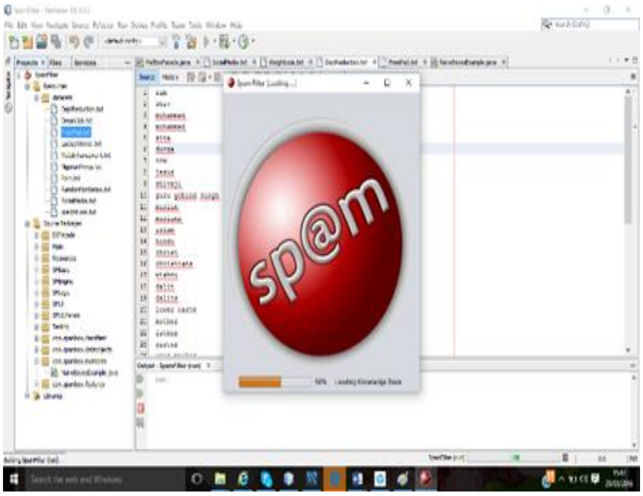


Fig. 4:

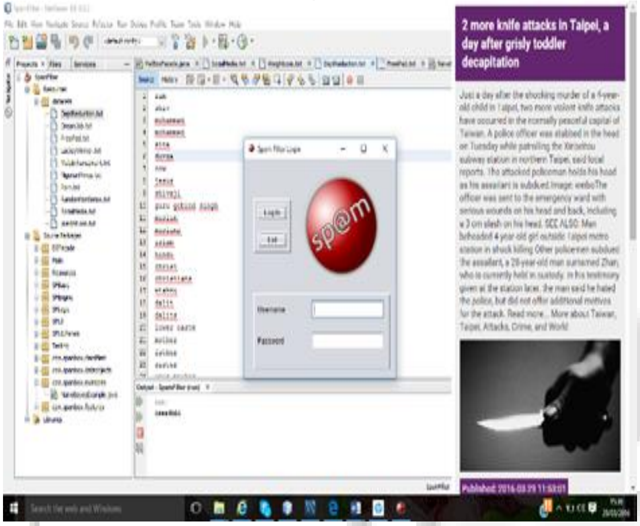

Spam Filter Login
Login
Exit
sp@m
Username
Password
2 more knife attacks in Taipei, a day after grisly toddler decapitation

Fig. 5:

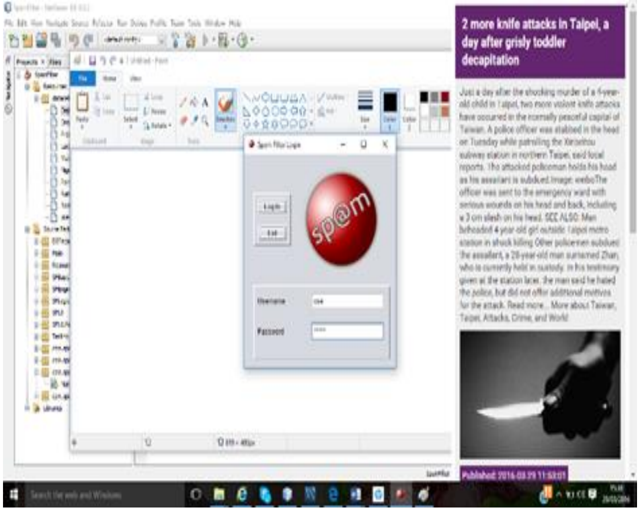

Spam Filter Login
Login
Exit
sp@m
Username
Password
2 more knife attacks in Taipei, a day after grisly toddler decapitation

Fig. 6:

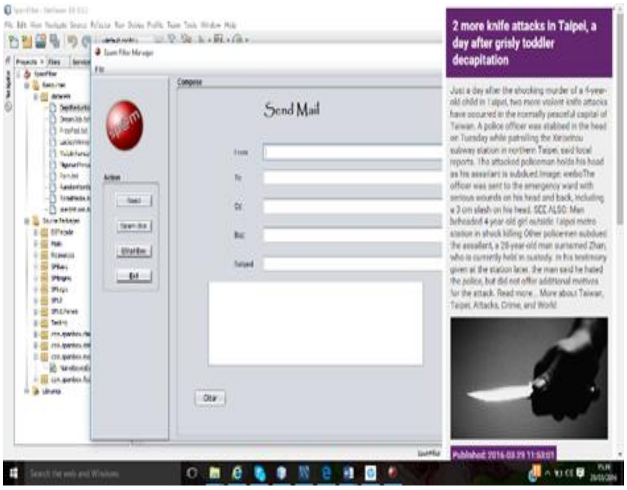


Fig. 7:

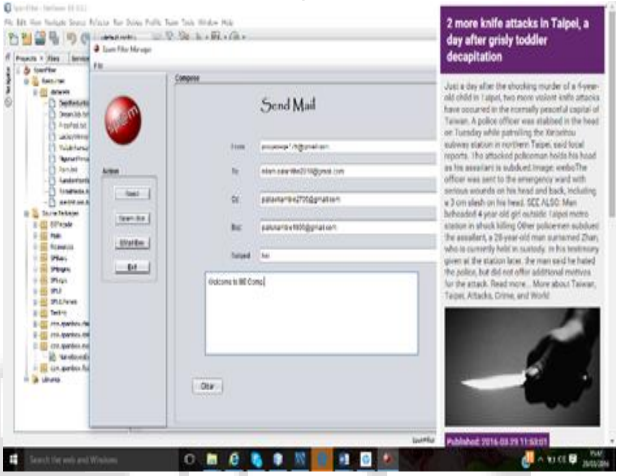


Fig. 8:

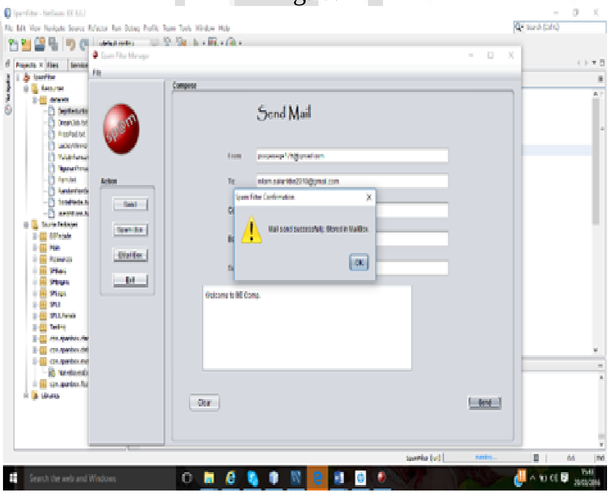


Fig. 9:

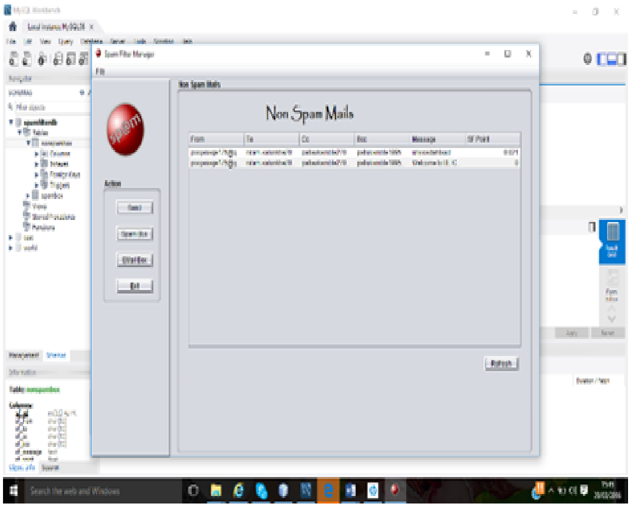


Fig. 10

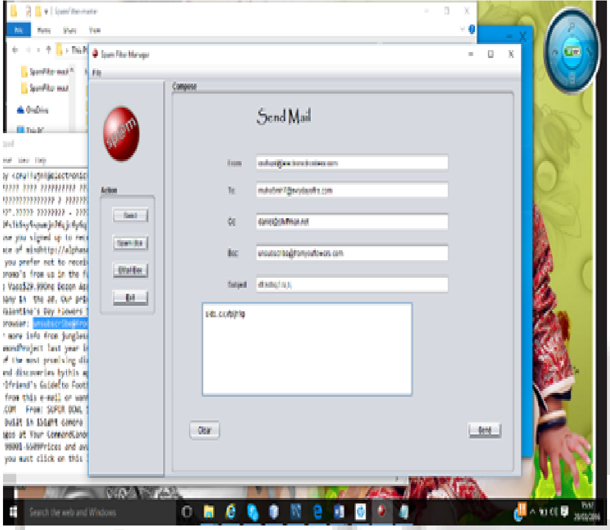


Fig. 11:

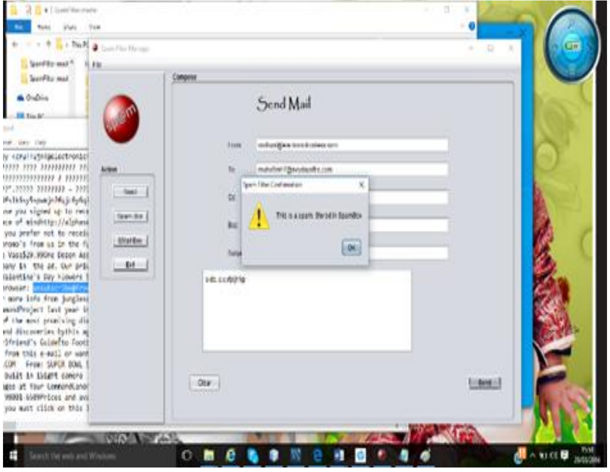


Fig. 12:

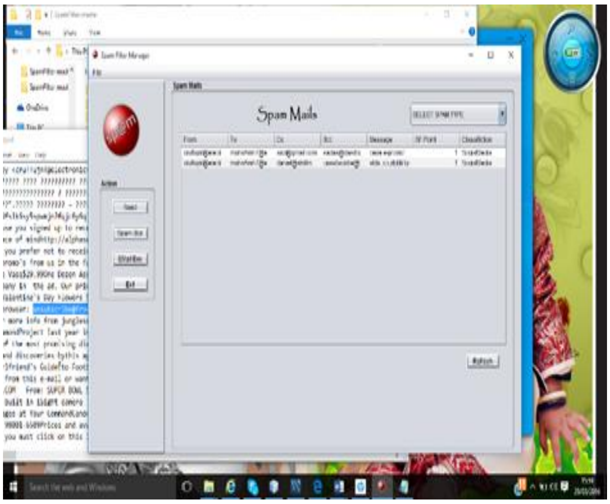


Fig. 13:

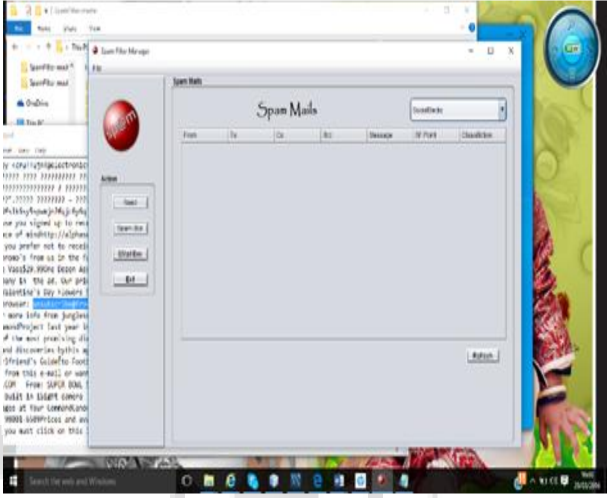


Fig. 14:

V. CONCLUSION

In this paper, an naive bayes approach is proposed to classifyspam emails. Extensive experiments conducted on a public spam email dataset indicate that the proposed algorithm outperforms the popular classification techniques including NaiveBayes. Our future work is to block the spam email and also unblock it.

ACKNOWLEDGMENT

The project topic would not have seen the light of the day without the whole-hearted support of my guide Prof. K. S. Warke. We admire her infinite patience and understanding that she guided me in a field We had no previous experience. Mam also guided us through the essence of time management, presentation skills and how vital it is for an engineer to think from a research perspective. Whenever We approached her, she explained the concepts lucidly, so that it would be simplified and be vivid in our mind.Again, we also thank to her and all the faculty members who have made the journey of our faculty directions in this domain. We thank to all of ourfriends and teachers, who have attended our seminar sincerely.

REFERENCES

[1] Global Journal of Computer Science and Technology Software & Data Engineering Volume 12 Issue 13

Version 1.0 Year 2012An Approach to Email Classification Using Bayesian Theorem By DenilVira, Pradeep Raja &ShidharthGadaK.J.Somaiya College of Engineering, Mumbai, India.

[2] IOSR Journal of Computer Science (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727 PP 68-72Effective Spam Detection Method for Email Savita Teli1, Santoshkumar Biradar2

[3] Spam Detection and Filtering using Different MethodsBhawanaS.Dakhare, V.Gaikwad Assistant Professor TernaEngg.College, Nerul, Navi Mumbai Mumbai,MEDHA-2012

[4] International Journal of Computer Applications (0975 – 888) Volume 47– No.16, June 2012 26 Spam Mail Filtering Technique using Different Decision Tree Classifiers through Data Mining Approach - A Comparative Performance Analysis SaritChakraborty Bikromadittya

[5] International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011 MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION W.A. Awad1 and S.M. ELseuofi2 1Math.&Comp.Sci.Dept., Science faculty, Port Said University

[6] International Journal of Innovative Research in Computer and Communication Engineering (An Vol. 3, Issue 4, April 2015Survey Paper on Effective Email Classification into Spam and Non-Spam Mails

[7] American Journal of Engineering Research (AJER) 2013 An Efficient Spam Filtering Techniques for Email Account S. Roy, A. Patra, S.Sau, K.Mandal, S. Kunar

[8] Anti-Spam Methodologies: A Comparative Study SaimaHasib, MahakMotwani, Amit Saxena