ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

Harmful Mail Scanning and Spam Filtering Through Data Mining Approach

Deepika Mallampati

Assistant Professor, Dept. of CSE, Sreyas Institute of Engineering and Technology, Nagole,Telangana,India

ABSTRACT: The increasing volume of unsolicited bulk e-mail (also known as spam) has generated a need for reliable anti-spam filters. Machine learning techniques now days used to automatically filter the spam e-mail in avery successful rate. A framework is used for evaluation of classifier security that formalizes and generalizes the training and testing datasets. We present novel method to access Classifier Security against their attack while presenting we are going to build SPAM filtering application with the help of "Bag Of Words Method" also we analysed the overall SPAM traffic on our Email Server.This paper discusses about the popular statistical spam filtering process naive Bayes classification.

KEYWORDS: Spam filtering,Machine learning algorithms, e-mail filtering

I. INTRODUCTION

In last few years, the increasing use of e-mail has resulted in the looks and additional acceleration of problems because of unsolicited bulk e-mail messages, typically known as to as spam. Evolving from a small irritation to a foremost problem, given the excessive circulating volume and no longer appropriate content material of a few of these messages, spam is to reduce the reliability of e-mail. Personal customers and organizations are plagued by unsolicited mail with admire to the use of network bandwidth utilized receiving these messages and the time wasted by customers classifying between junk mail and common (respectable or ham) messages. A businessmodel depending on unsolicited mail marketing is mainly outstanding due to the fact the costs for the sender are much less, it tends tremendous number of messages may also be sent, this aggressive behaviour being probably the most primary characteristics of Spammers. One other method got is the usage of junk mail filters, which is based on gain knowledge of the message contents and additional know-how, effort to classify junk mail messages. The action to be in use once they are identified by and large is determined by the surroundings where. If used as a clientside filter it is embed nearby on client approach and classify mails labelled as spam of legimate. Whereas a further is server aspect filer it is reward on mail serves supplier server dealing with more than a few messages as junk mail and send to respective user.

Skills engineering and computer studying are the 2 common strategies used in e-mail filtering. In potential engineering process a set of rules has to be distinctive in line with which emails are classified as spam or ham. A suite of such rules must be created either by using the consumer of the filter, or by means of any other authority (e.g. the software corporation that provides distinct rule-founded spam-filtering software). With the aid of making use of this approach, no promising results indicates on the grounds that the principles have to be always up-to-date and maintained, which is a waste of time and it isn't effortless for most clients.

Typically e-mail is split into three constituents Header, Subject line, Bodywhereas Header involves the information of sender, receiver and so forth. Subject Line. In which discipline the e-mail is and last subject is content or physique of e-mail this is exact content material of email. Earlier than the available expertise can be utilized by a classifier in a filter, correct pre-processing steps are required. The steps worried within the extraction of information from a message are illustrated in fig.1 beneath and will also be grouped into:

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

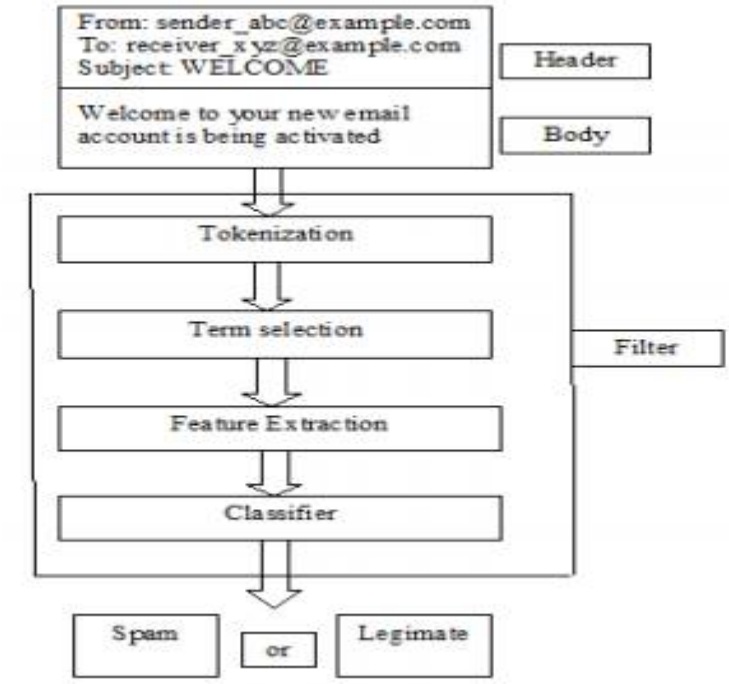

Fig.1: Architecture of spam filter

(1) Tokenization, in which the words get extracted from the message body.
(2) Term Selection, which rank each term according to the irusefulness.
(3) Feature Extraction, which gives reduced set of data

In this paper we are using document frequency as our tern selection method whereas Bag of words is our feature extraction based on local feature extraction and Naive Bayes classifiers of spam and legitmate respectively.The paper is organized as follows: section 1 is the paper introduction, section 2 summarize the related work done using various algorithms, Classifiers inEmail Spam Filtering section 3 gives a general theoretical description oh proposed method we used in this study, section 4 present detailed steps of the experiment implementation and performance comparison of the methods, finally we closed the paper with the conclusion insection 5.

II. RELATED WORK

In the study [1] Rambow et al. applied machine learning techniques for email summarization. In this study, RIPPER classifier is used for the resolve of sentences which should be included in a summary. Learning model use features such as Linguistic feature, email features, and threading structure. This approach entails positive examples in huge quantity and it is also found that summaries are not produced for varying length based on user interest.

There are such a lot of current strategies for detection of these spam emails. These techniques come ordinarily from the discipline of artificial Intelligence, data Mining, or Machine learning. Machine learning schemes are more assorted and used broadly for spam mail classification. Selection tree classify spam mails making use of prior data [2. But it is highly-priced to calculate and recalculate as spammers alternate procedure. In the literature [3] Bayesian networks located as the very preferred technique for junk mail detection. However with this strategy it is quiet complicated to scale up on many elements to come out with the judgement.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

In [4] fuzzy clustering procedure is used. On this paper author evaluated the usage of fuzzy clustering and textual content mining for spam filtering. Fuzzy clustering is scalable and effortless to update process. This is trained offers with the examination of use of fuzzy clustering algorithm to construct a spam mail filter. Classifier has been proven on one-of-a-kind data units and after testing Fuzzy C-approach making use of Heterogeneous value difference Metric with variable percentages of spam mail and used a regular model of assessment for the hindrance of spam mail classification.

This paper makes use of textual content mining and fuzzy clustering as an anti-spam mail process. If every electronic mail that comes in is used as part of the data pool to make choices about future emails, spam trends might be detected. It is determined that there is not huge price of calculation and recalculation that might arise with selection tree, or with some rule-based filters.

In the paper [5] two ways are described for classification. First is done with some rules which can be defined manually, like rule headquartered trained method. This process of classification is utilized when lessons are static, and their add-ons are simply separated in accordance with the features. Second is completed with the support of present computing device learning methods. According to the be trained [6] clusters of spam emails are created with the help of criterion perform. Criterion function is outlined because the maximization of similarity between messages in clusters and this similarity is calculated utilising ok-nearest neighbour algorithm.

Symbiotic data Mining is a distributed data mining strategy which unifies content founded filtering with collaborative filtering is described in [7]. The major goal is to make use of local filters once more with a view to support customized filtering in context of privateness. In study [8] email classifiers cantered on the method of feed ahead again propagation neural community and Bayesian classifiers are evaluated.

In the paper [9] Bayesian approach is applied for the problem of classification and clustering using model based on the assumptions like: population, subject, latent variable, and sampling scheme.

In the paper [11] spam is detected using artificial neural network. In this paper author designed the artificial neural network spam detector using the perceptron learning rule. Perceptron employs a stochastic gradient method for training, where the true gradient is evaluated on a single training example and the weights are adjusted accordingly until a stopping criterion is met.

III. PROPOSED METHODOLOGY

In this work we control disorders above by way of utilising constructing up a structure for the observational assessment of classifier protection at configuration stage that extends the model separate and execution analysis schemes of the based define cycle .We compress front work, and speak to awareness to a couple predominant originations that rise up out of it. We then formalize and sum them up in our procedure.

1) To start with, to go looking after defense in the connection of a weapons contest it is not adequate to reply to located attacks, nonetheless as a substitute it is usually obligatory to proactively suspect the adversary by way of guessing typically probably the most suitable, capabilities attacks via an count on a situation the place investigation; this authorizations one to create suitable counter measures upfront of the assailment truly occurs, as per the guideline of protection via configuration.

2) To give practical rules to recreating particular assault situations, we signify a usual model of the enemy, related to her purpose, discernment, and skills, which include and sum up units proposed in fundamental work.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

3) On the grounds that the neighbourhood of scrupulously involved with attacks may just influence the conveyance of constructing capable and trying out expertise discretely, we suggest an datamodeldistribution that may formally describe this comportment, and that authorizes us to recollect a vastly huge number of talents attacks; we withal advocate a calculation for the science of making in a position and checking out models to be used for safeguard evaluation, to be able to relatively quite often swim go well with software-concrete and heuristic systems for mimicking attacks.

Relieson very simple representation on document

- Bag of words

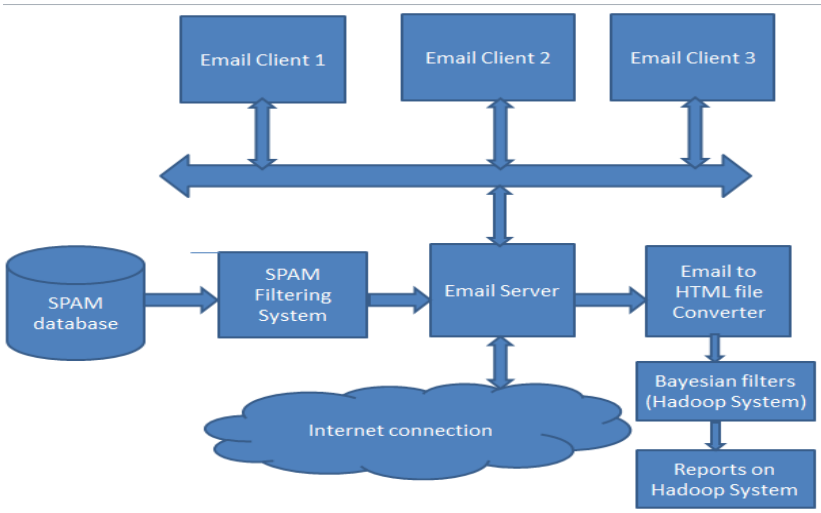

Fig.2: Proposed system

Bag of word model is one in all the widely used feature extraction method in spam filtering applications. It converts a message to a d-dimensional vector [x_1, x_2….. x_d] by considering occurrence of already selected terms that are chosen by utilizing term selection method. Within the vector, x_i will be viewed as a function of the term ti's occurrence within the message.

A. Naive Bayes classifiers

Naive Bayes classifiers are a mostly used technique of e-mail filtering. This method involve bag of words features to classify spam e-mail, this technique commonly used in text classification. Naive Bayes classifiers use Bayesian assumption to calculate a probability that an email is or is not spam. In Bayesian framework the messages get divided in to different representations then the probability that given representation of message, denoted as representation of a message, denoted as x'=(x_1,x_2,x_3,….,x_n),belongs to class c is given by:

$P(c/x) = [P(x/c)P(c)]/(P(x))$
Where,
P(c/x) is probability of term to be occur in message
P(x) is probability of term to be occur in message

This probability can be used to make the decision about if a message should belong to the category. In order to compute this probability, estimations about the message in the training set are made. When computing probabilities for spam classification we can prevent the denominator because we only have 2 classes (spam and legitimate), and one message

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

cannot be classified into only one of them, so denominator is the same for every set of words can be estimated as the number of message in the training set belonging to the category, divided by the total number of documents.

Calculating P(x/c) is a bit more complicated because we need in the training set some messages identical to the one we want to classify. When using Bayesian algorithm it is very frequent to find the assumption that terms in a message are independent and the order they appear in the message is irrelevant. This way probability can be calculated as

$$p\left(\frac{x}{c}\right) = \prod_{i=1}^{y} p\left(\frac{ti}{c}\right)$$

The statistic we are mostly interested for a token T is its spamminess (spam rating) [10], calculated as follows:

$$S[T] = \frac{C_{spam}(T)}{C_{spam}(T) + C_{ham,}(T)}$$

Where $C_{spam}(T)$ and $C_{ham,}(T)$ are the number of spam or ham messages containing token T, respectively. To calculate the possibility for a message M with tokens $\{T_1,.....,T_N\}$, 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

$$(H[M] = \prod_{I=1}^{N} (1 - S[T_I]))$$

The message is considered spam if the overall spamminess product S[M] is larger than the hamminess product H[M]. The above description is used in the following algorithm [10]:

Stage1. Training

Parse each email into its constituent tokens
Generate a probability for each token W
S [W] = $C_{spam}(W)$ / $C_{spam}(W)$ + $C_{ham,}(W)$store spamminess values to a database

Stage2. Filtering

For each message M
While (M not end) doscan 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 asI [M] = 1+S [M]-H [M]
if I [M] > threshold
msg is marked as spam
else
msg is marked as non-spam

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

As shown in fig.3 the complete information of work flow of our spam filtering is explained with the help of sequence diagram.

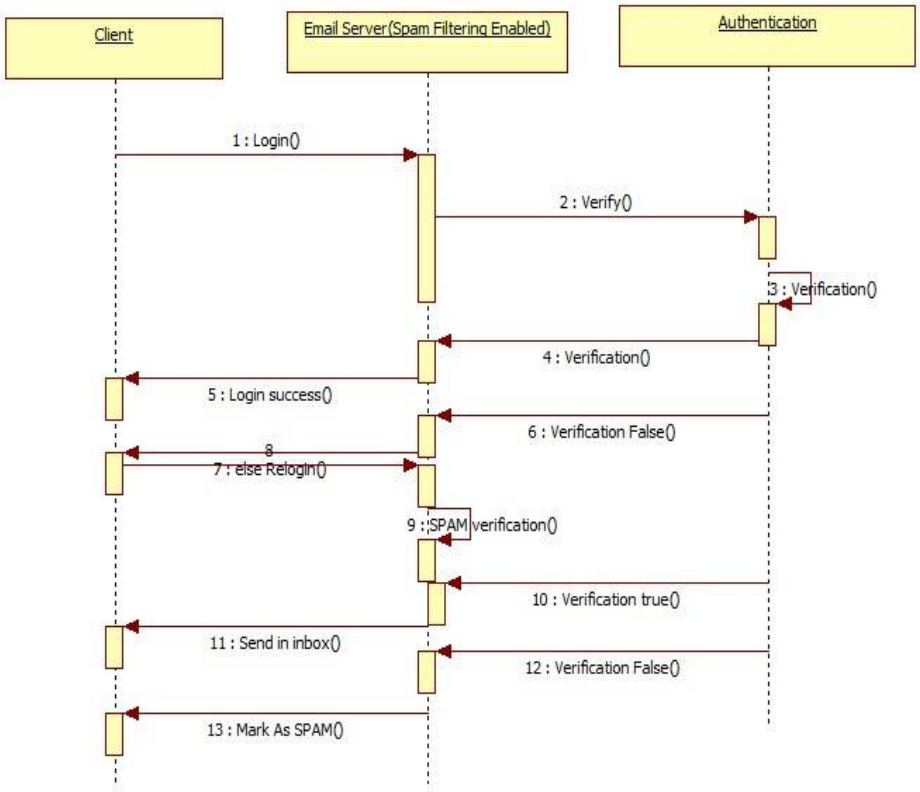

Fig.3 Sequence diagram of proposed system

IV. RESULTS

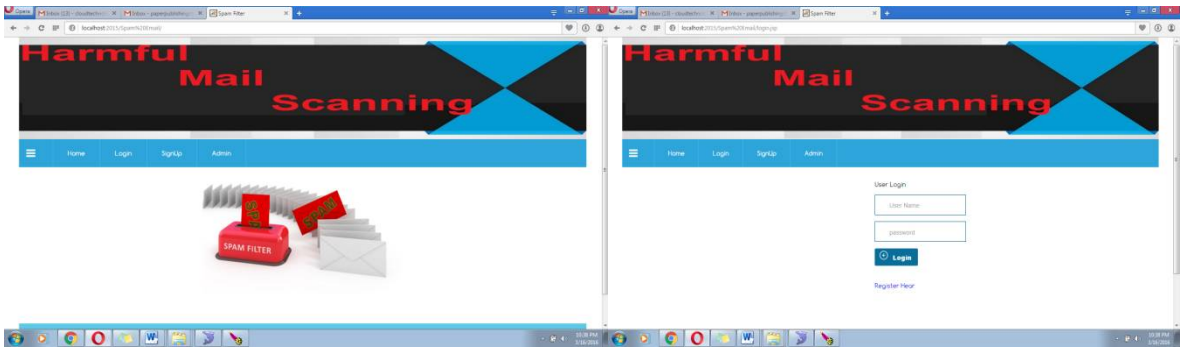

Fig.4: Spam filter home screen

Fig.5: User Login

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

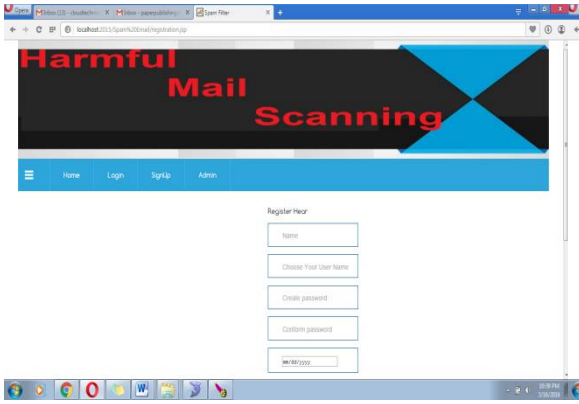

Fig.6: Register Login

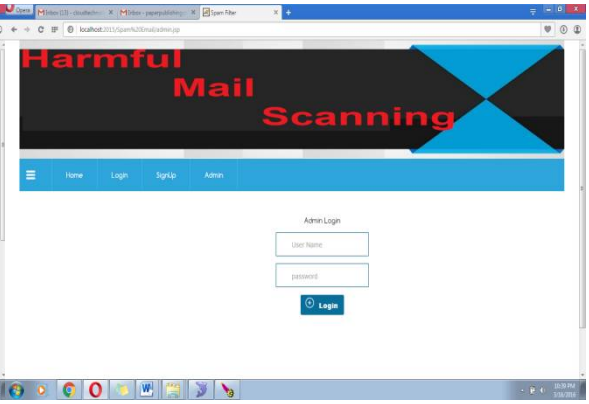

Fig.7: Admin Login

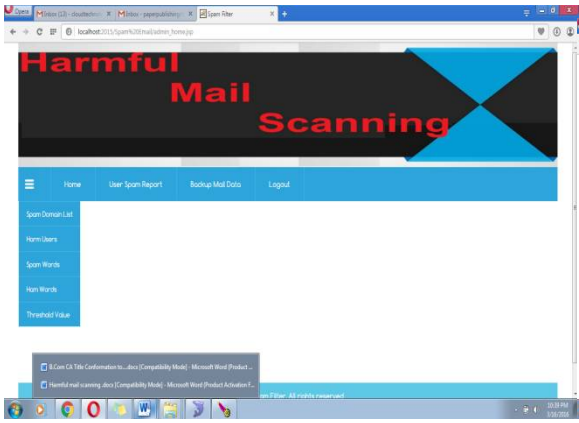

Fig.8: User View

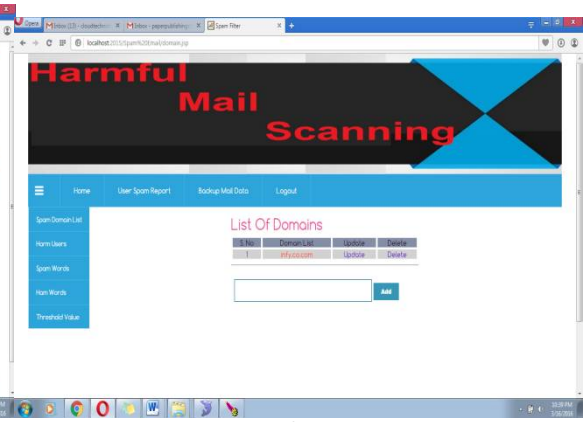

Fig.9: List of Domains

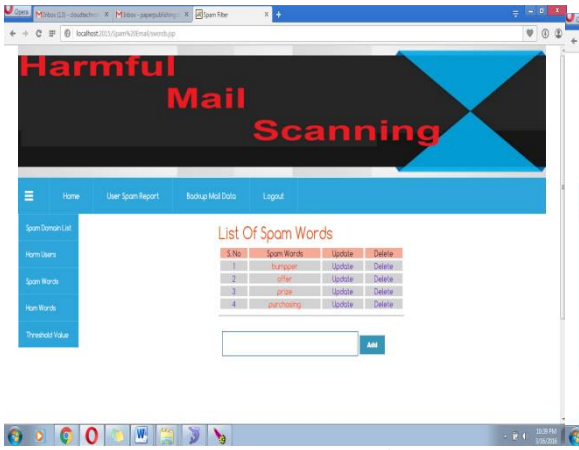

Fig.10: List of Spam words

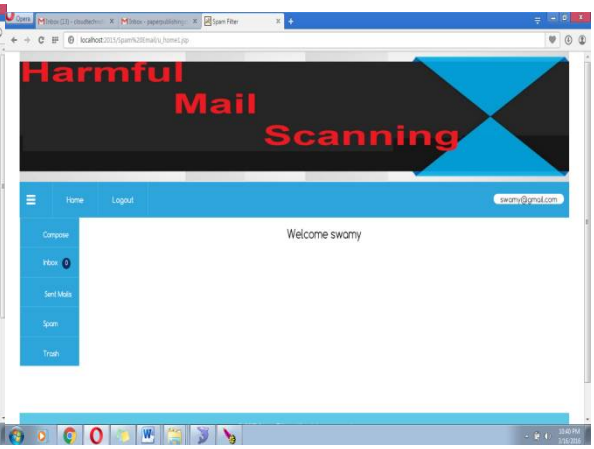

Fig.11: Welcome Screen

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

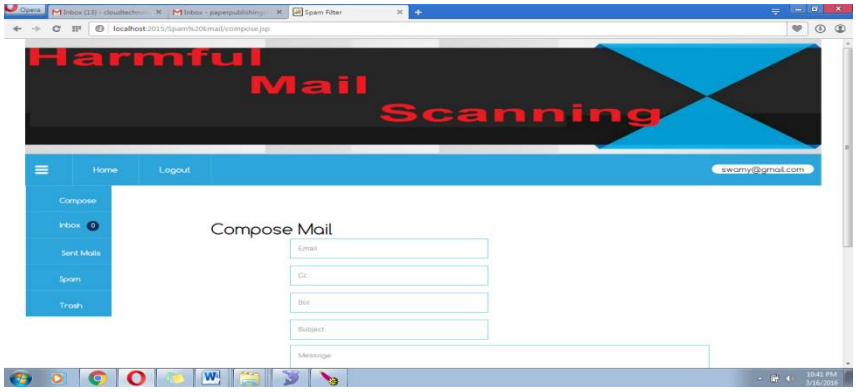

Fig.12 Mail compose

V. CONCLUSION

In this paper we review some of the most popular spam filtering methods and of their applicability to the problem of spam e-mail classification.We observed Naïve bayes has a very satisfying performance among the other methods, more research has to be done to escalate the performance of the Naive bayes.This method currently detects spam only in text but can further accommodate other features like images, video and social network features as well. The complexity of this approach is low and it can be used in reality easily.Our work prevents developing novel methods to assess classifier security against these attacks. Finally the presence of an intelligent and adaptive adversary makes the classification problem highly non-stationary.

REFERENCES

[1] Ducheneaut N and Bellotti V. E-mail as habitat: anexploration of embedded personal informationmanagement [A]. Interactions ACM, 2001, 8: 30- 38.
[2] Carreras X, and Marquez L. Boosting trees for antispam filtering [C]. In International conference onRecent Advances in Natural Language Processing. , 2001 160-167.
[3] Sahami M, Dumasi S, Heckerman D, and HorvitzE. A Bayesian approach to filtering junk e-mail: InLearning for text categorization [A]. Papers fromthe 1998 Workshop, Madison, Wisconsin, 1998.
[4] Mohammad N.T.A Fuzzy clustering approach tofilter spam E-mail [A].Proceedings of WorldCongress on Engineering, vol. 3, WCE-2011.
[5] Biro I, Szabo J, Benczur A, and Siklosi D. LinkedLatent Dirichlet Allocation in Web Spam Filtering[A].In Proceedings of the 4th InternationalWorkshop on Adversarial Information Retrieval onthe Web (AIR Web), Madrid, Spain, 2009.
[6] Perkins A. The classification of search engine spam.http://www.ebrand management.com/whitepapers/spam classification, 2001.
[7] Paulo C, Clotilde L, Pedro S. Symniotic datamining for personalized spam filtering [C]. In theWeb Intelligence and Intelligent Agent Technology,2009, 149-156.
[8] Rasim M A, Ramiz M A, and Saadat A N.Classification of Textual E-mail spam using DataMining Techniques [J]. In the Journal of AppliedComputational Intelligence and Soft Computing,2011.
[9] Erosheva E A and Fienberg S E. Bayesian mixed membership models for soft clustering and classification [J]. Proceedings of National Academy of Sciences, 2004, 97(22):11885-11892.
[10] Li, K. and Zhong, Z., "Fast statistical spam filter by approximate classifications", In Proceedingsof the Joint international Conference on Measurement andModeling of Computer Systems. SaintMalo, France, 2006
[11] Kufandirimbwa O, Gotora R. Spam detection usingArtificial Neural Networks [J]. In Online Journal ofPhysical and Environmental Science Research,2012, 1:22-29.

BIO DATA

Deepika Mallampatiattained her B.Tech in Computer science and Information Technology and M.Tech in the stream of Software Engineering from JNTU, Hyderabad. She is having teaching experience of more than 8 years in various Under Graduate and Post Graduate course. She has guided lots of students in various Under Graduate and Post Graduate Research Projects. At present, she is working as Assistant Professor, Department of CSE in Sreyas Institute of Engineering And Technology, Nagole, Hyderabad, Telangana, India