RESEARCH ARTICLE

Automated Spam Filtering through Data Mining Approach

*Deepika Mallampati[1], Amitesh Mathur[2], Gundari Abhinay[2], Gopalam Tanuja[2]
[1]Assistant Professor, Department of Computer Science & Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, Telengana, India.
[2]Department of Computer Science & Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, Telengana, India.

Received- 18 November 2016, Revised- 26 December 2016, Accepted- 9 January 2017, Published- 29 January 2017

ABSTRACT

Spam messages can be referred as those mails which come into act in the absence of a standard agreement among the senders and receivers for receiving e-mail solicitation. Usually these messages are sent in bulk quantities. For preventing the spam delivery, an automatic system based spam filter tool is employed. The objectives of spam filters and spam are contradicted diametrically. A spam filter can be termed effective if it recognizes spam. On the other hand, it is ineffective when it escapes the filters. It is the need of the hour that these bulk unsolicited e-mails be effectively filtered. Increasing volume of these mails emphasizes on the requirement and design of dependable anti-spam filters. One of the techniques which is used widely to filter these spam e-mails is the machine learning technique. They possess in built algorithms which filters spam e-mails at commendable rates. In this project we present a method, to access classifier security against their attacks profoundly concentrating on the content of the message. The dependence on a predefined set of keywords is reduced. The paper also focuses on related works which apply machine learning techniques using naïve Bayes classification for e-mail message classification.

Keywords: E-mail, Spam, Spam filtering, E-mail classification, Feature extraction.

1. INTRODUCTION

The paper is organized as follows: Section 1 explains the basic concepts. Section 2 briefs the different literatures available, and touches different algorithms and classifiers in e-mail spam filtering. Section 3 provides a specific theoretical explanation of the method proposed. Section 4 analyses the elaborate steps used for the implementation of the project. It also compares the earlier available methods and their result to that of the obtained results. Section 5 describes the conclusion and the future scope of the proposed project.

The occurrence and threats of e-mail spams has increased at a fast pace with the advent of technology. Users, both business as well as normal household users and network administrators are more concerned with this increase [2]. Statistics as of July 1997 clearly indicates that spam messages form approximately 10% of the incoming messages while considering a corporate business network.

Further, Message Labs [1], in its 2006 Annual Security Report stated that spam messages and related activities has shown a steady increase particularly in 2006. They showed that such an increase is equivalent to 86.2% of the whole e-mail traffic. Another indication was that with the increased advent and availability of sophisticated robotic technology, a.k.a.botnets the spam volumes increased by 70% in the later half of 2006. This inturn has also added to the increase in the overall email traffic. Taking into account the current trends and by analyzing, it was feared that spam deliverance will continue to rise and reach the peak at around 92% of e-mail traffic by the last quarter of 2007. It also predicted that by the end of 2015, spam

*Corresponding author. Tel.:
Email address: deepika.m@sreyas.ac.in (D.Mallampati)
https://dx.doi.org/10.24951/sreyasijst.org/2017021004
Double blind peer review under responsibility of Sreyas Publications

messages will go beyond 95% of the total e-mail traffic. Eventhough the accuracy of these figures is questioned, it can be right way inferred that spam volume is dramatically on the rise over the years.

Advancements in internet technology has resulted in new channels of communication. E-mail, has the primary advantage of senting a mail to a relative who is thousands of kilometres away. This makes e-mail vulnerable for mass e-mailing, reaching out to hundreds of thousands of users with in a fraction of a second. However, as in the case of most other technologies, the freedom is misused here too.

Spam can be destructive to the recepients in a multitude of senses. According to Ferris Research, if an employee receives, five e-mails per day and consumes 30 seconds on each, then he/she is prone to waste 15 hours of valuable work time in an year. Multiply this with the employee hourly rate will show the loss inurred by the company by means of spam messages. Another potential area of concern is that spam softwares can be used to spread and replicate harmful web contents such as Trojans, viruses, worms and different malicious codes. Phishing attacks are also a necessary consequence of spams.

All these factors has led to the notion that spam is an area which requires requires great attention, thereby motivating researchers and practitioners in finding novel techniques to curb them. Additionally with the improved legislations and regulations, different anti-spam technical solutions have been put forward. These techniques were then executed to combat the problem. By studying such novelties, it is found that most of them are static which means that they made use of a blacklist spammer, a good source white list or a fixed keyword set for identification of spam messages. Although the risks were reduced substantially, with new techniques scaled and adapted by the tactics of spammers, the efficiency of such techniques also got less efficient. Some of the spammers used techniques such as changing the address every time, intentional misspelling of words, content forging etc., to get rid of the designed bypass spam filters.

1.1. Definition

The term SPAM refers to unsolicited, unwanted, inappropriate bulk email. SPAM is the abbreviated form of Stupid Pointless Annoying Message.

1.2. Types of spam

Based upon the source spam has different definitions.

•Unsolicited bulk e-mail (UBE)—unsolicited e-mail, sent in large quantities.

•Unsolicited commercial e-mail (UCE)—this more restrictive definition is used by regulators whose mandate is to regulate commerce. [Any email message that is Fraudulent].

•An email message where the sender's identity is forged, or messages sent though Unprotected SMTP servers, unauthorized proxies.

1.3. Characteristics of spam

The spam characteristics are presented in two different parts of a message: message content and e-mail headers

E-mail headers: E-mail headers display the path an e-mail has taken for arriving at the destination. They also contain other information connected to the e-mail, such as intended receiver and sender, the ID of the message, the transmission date and time, subject and several other mail characteristics.

1.4. Typical email header characteristics in spam messages

- Recipients email address is not in the To: or Cc field
- Empty To: field
- To: contains the invalid email address
- Missing To: field
- From: field is same as To: field
- Missing From: field
- More than 10 recipients in To: and/or Cc: fields
- Bcc: header exists
- Message contents

The email filtering system should filter out spam messages in three ways:

- Block spam at the gateway itself by checking domains in real time black hole lists.
- Filter out spam based on email characteristics
- Identify Junk mail content

2. RELATED WORK

2.1. Techniques to eliminate SPAM

Based on the survey performed, this section summarizes, some common techniques for preventing spam and shortly describes the different spam filtering techniques which is presently in use.

2.2. Hiding the e-mail address

The most practical approach for spam avoidance is to fix the e-mail address in such a way that it is not accessed by the spammers. Only trusted parties should be provided with the e-mail address. Temporary e-mail accounts can be used for less trusted parties. In the case of displaying the e-mail address on the website, it can be disguised for e-mail spiders by means of a tag insertion that is requested to be removed before replying.

Specialized robotic entities will collect the tagged e-mail address. On the other hand, a human would easily understand that the removal of tag is mandatory for retrieving the exact e-mail address. For many of the customers, this method is insufficient. Firstly, it is very time consuming act for employing techniques which will keep the e-mail addresses under safe custody. Secondly, eventhough the robots could be misled by the disguised address it could also mislead an inattentive human. However the protection remains, only till the time the e-mail address is exposed.

2.3. Pattern matching, white lists and blacklists

This type involves a content-based pattern matching approach, in which the incoming mails is correlated against pre-available patterns and are classified as legitimate or spam. Most of the e-mail programs have the feature which is better known as “message filters” or “message rules” which does this function. The technique is mostly carried out by means of plain string matching. Blacklists and whitelists, which denotes the list of friends and non-desired persons respectively comes under this category. When an incoming mail closely matches a white list entry, the e-mail will be allowed through the rule.

On the contrary, if the incoming mail matches the blacklist it will be treated as spam. Spams are reduced to a considerable level by this method. However constant updating is required in this case, as spams constantly evolve. It is very time consuming to find which rules are apt for spam removal and spam identification is not good with this technique. Eventhough the results claim that 80% of spams were able to be caught, it had very high false positive rates. Technically, the above mentioned method is a basic version of the highly improved“rule based filter” which are discussed below.

2.4. Rule based filters

This is one of the popular content based method [3, 4] employed by using spam filtering softwares such as Spam Assassin. In this method, each incoming mail is treated with a set of rulesby the rule-based filters. In the case of a match, a score is assigned to each mail which indicates non-spaminess or spaminess. If the total score crosses the threshold score, the e-mail will be automatically classified as a spam. Regular expressions make up the rules, which is accompanied by the software. In this case too, update of the rule set must be done at regular intervals. This enables the enhanced successful rates of spam filtering. Updates are done through the internet. The comparison test results of ant-spam programs depicts that Spam Assassin recognizes 80% of all the spams. However the more improved statistical filters, were able to find 99% of all spam.

The rule-based filters have the advantage that they do not require any training for reasonable performance enhancement. Humans implement these complex rules. Before implementing a newly proposed rule, testing should be carried out extensively, inorder to make sure that, only spam messages are designated as spam and legitimate messages are not treated as spam. Similar to other methods discussed above this method also requires frequent rule updation. Once the spammer finds a mode to deceive the filter, with the same set of rules, the spam messages will get rid of all filters.

2.5. Statistical filters

Remarkable results are possible by employing a statistical spam classifier [7, 8]. Over the course of time many statistical filters have evolved; the reason being simplicity, ease of implementation, performance guarantee, and low maintenance cost. Training is an integral part of these statistical filters and gradually they will become more and more efficient. They are trained personally on the legitimate and spam e-mails of the user. Using this technique it is very hard for the spammer to deceive the filter.

2.6. E-mail verification

E-mail verification refers to a challenge-response system which normally sends a one-time verification e-mail to the sender. Only if the sender successfully responds to the challenge the e-mail will pass through the filter. The challenge is usually by means of a hyperlink send in the form of a verification mail which the sender is required to click. Once this link is clicked, e-mails from the sender are allowed. Choice mail and blue bottle are two such systems. This method is able to filter almost 100% spam. However, this method has two drawbacks.

The sender needs to respond to the challenge which consumes more time and requires extra care. Once the challenge is not fulfilled the e-mail will be lost.

Verification e-mails can also be lost because of technical obstacles, similar to fire walls and different automatic e-mail response systems. It can also result in problems for automated e-mail responses such as newsletters and online orders. More traffic is generated by the verification mail which is also a disadvantage.

2.7. Distributed blacklists of spam sources

Here the filters employ a distributed backlist [3] to find whether an incomimg e-mail is a spam or not. The internet is the source of the distributed blacklist the users of the filter constantly updates the distributed blacklist. If a spam passes through a filter, the blacklist is updated. Now the users will be protected from the sender of that particular e-mail. This blacklists class keeps the known spam sources as a record in the form of IP numbers that allow SMTP relaying. However, it also has some disadvantages by depending entirely on the blacklists. The false positive approaches hinders the entire output. The time consumption for the network based loop is another downside. These solutions may be useful for companies assuming that all their e-mail communications are with other serious non-listed businesses.

2.8. Distributed blacklist of spam signatures

The major difference [3] between this method and the previously described method is that the blacklists are made up of spam message signatures instead of spam sources. When a spam message is received by the user, the same user can report the message signature(typically a hash code of the e-mail) to the blacklist. In this process, a particular user will be able to warn other potential users about the authenticity of the message. For avoiding non-spam addition to a blacklist, multiple users must have reported the same signature. However the spammers invented an innovative way for fooling these filters; it is by means of adding a random string to each and every spam, thereby preventing the detection in the black list. Another disadvantage is the time taken for the network lookup.

Countering this measure, spam fighters overcome it, by including random noise by means of their signature algorithms. The legitimate messages are thereby rarely classified as spam. Most of the spams are also not recalled. Vipul’s Razor makes use of such a blacklist and states that it finds 60%-90% of all incoming spam.

2.9. Fuzzy clustering method

In [9, 10] fuzzy clustering procedure is used. In this paper the author analyzed the fuzzy clustering usage and mining of textual content for spam filtering. For updating the process fuzzy clustering is effortless and scalable. This is trained 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.

2.10. Bayesian Classification

In [8, 9], for the problem of clustering and classification, Bayesian approach is applied. The classification is based on assumptions like subject, population, sampling scheme and latent variable

2.11. Artificial neural network

In [5, 6, 7] artificial neural networks are employed to detect spam. Here perceptron learning rule, is used to design the artificial neural network which means that a stochastic gradient method is used for training and the true gradient is analysed on a sigle training example. Till the achievement of the stopping criterion the weights are adjusted.

3. PROPOSED METHODOLOGY AND EXPERIMENTAL RESULTS

In this paper, we a use document frequency as our term selection method where feature extraction is done based on local feature extraction and Naive Bayes classifiers of spam and ham respectively

The classification of e-mail tasks are primarily understood by means of several subtasks.

E-mail classification tasks are often divided into several sub-tasks. The first step is the data collection and representation. Both of them are problem specific (i.e. e-mail messages). The second step is the e-mail preprocessing, feature extraction and feature reduction. The feature reduction process helps in reducing the dimensionality (i.e. the number of features) for the remaining steps of the task. The final step is the e-mail classification step finds the original mapping between the training and testing sets [11-14].

Figure A1 and A2 describes the working model of the proposed system and the sequence diagram of the proposed system. Figures A3 to A14 shows the obtained results.

3.1. Detailed algorithm steps

Step 1: Email preprocessing

In step 1, the corresponding stop words, i.e. words that frequently appear but have less power of discrimination are taken from the e-mail message. ‘a’, ‘an’, ‘and’, ‘the’, ‘that’, ‘it’, ‘he’, ‘she’…etc. are examples of such words. Next the message is tokenized into set of strings separated by delimiters e.g. white spaces. Tokens can symbolize phrases, word or any key word patterns. Next the lower case conversion of mixed-case tokens take place. There is a large difference between treating uppercase and lower case letters. The set of resulting tokens are stemmed to their roots. This avoids considering different forms of the same word as different attributes. This, considerably reduces the attribute set size. Now all message tokens are combined into one vector T=<t1, t2…, tN> where N is the total number of tokens. Additionally, the frequency of each token occurences, in each category c (spam, ham) is determined.

Step 2: Training

Depending on each category characteristics, while training, a model is built in a set of pre classified e-mail messages. Subject and content should vary for each training data set. The ham and spam texts are extracted by the feature extraction module. It then produces the feature vectors and dictionary as the selected algorithm input for training and testing the classifier [9]. In the training part, this module accounts for the frequency of words in the email text. In our approach we take words whose time of appearance is more than three times as the feature word of the class. Every e-mail in training is denoted as a feature vector.

Step 3: Spam classification

After the above mentioned steps, we assign the standard classification email documents as training document. Next the following processes viz., e-mail pretreatment, extracting of useful information, save into text documents according to fix format, splitting the whole document to words, extracting the feature vector of spam document and translating into the form of vector of fix format etc., are carried out. We always concentrate on optimal classification by means of selected algorithms which is built by employing feature vectors of spam documents.

Step 4: Performance evaluation

For testing the performance of the methods mentioned above, we made use of the most popular evaluation methods used by the spam filtering researchers viz Spam Recall (SR), Spam Precision (SP) and Accuracy (A). Spam Precision (SP) is the number of relevant documents identified as a percentage of all documents identified; this shows the noise that filter presents to the user (i.e. how many of the messages classified as spam will actually be spam

$$\mathrm{SP} = \frac{\#\text{ of Spam Correctly Classified}}{\text{Total }\#\text{ of messages classifies as spam}} - \frac{N_{\mathrm{spam}\to\mathrm{spam}}}{N_{\mathrm{spam}\to\mathrm{spam}} + N_{\mathrm{ham}\to\mathrm{spam}}}$$

Spam Recall (SR) is the percentage of all spam emails that are correctly classified as spam.

$$\mathrm{SR} = \frac{\#\text{ of Spam Correctly Classified}}{\text{Total }\#\text{ of messages}}$$

$$= \frac{N_{spam\rightarrow spam}}{N_{spam\rightarrow spam} + N_{spam\rightarrow ham}}$$

Accuracy (A) is the percentage of all emails that are correctly categorized

$$A = \frac{\#\text{ of e} - \text{mails Categorized}}{\text{Total }\#\text{ of e} - \text{mails}}$$

$$= \frac{N_{ham\rightarrow ham} + N_{spam\rightarrow spam}}{N_{ham} + N_{spam}}$$

where $N_{ham\rightarrow ham}$ and $N_{spam\rightarrow spam}$ are the number of messages that have been correctly classified tothe legitimate email and spam email respectively. $N_{spam\rightarrow ham}$ and $N_{ham\rightarrow spam}$ are the number of legitimate and spam messages that have been misclassified; N_{ham} and N_{spam} are the total number of legitimate and spam messages to be classified. As shown in figure 2 the complete information of work flow of our spam filtering is explained with the help of sequence diagram.

5. CONCLUSION

Spam emails are the biggest problem for web data. This paper explored different approaches to deal withthis problem. We observed Naïve Bayes has a very satisfying performance compared to other methods. More research has to be done for performance escalation of Naive Bayes. This method currently detects text based spams but in future can further accommodate other features like social network features, images, video etc.,.

REFERENCES

[1] Saadat Nazirova, Survey on Spam Filtering Techniques, Communications and Network, Vol. 3, 2011, pp. 153-160, http://dx.doi.org/10.4236/cn.2011.33019

[2] Gary Robinson, A Statistical Approach to the Spam Problem, Linux Journal, 2003

[3] J.Rajeesh and E.Arun, Region Growing and Level Set Compound for Hippocampus Segmentation, DJ Journal of Advances in Electronics and Communication Engineering, Vol. 1, No. 1, 2015, pp. 1-7, http://dx.doi.org/10.18831/djece.org/2015011001

[4] Ahmed Khorsi, An Overview of Content- Based Spam Filtering Techniques, Informatica, Vol. 31, 2007, pp. 269-277.

[5] M.T.Banday and T.R.Jan, Effectiveness and Limitations of Statistical Spam Filters, International Conference on New Trends in Statistics and Optimization, Srinagar, 2009.

[6] M.Thangamani and P.Seetha Subha Priya, Image Retrieval System by Skin Colour and Edge Information, Journal of Excellence in Computer Science and Engineering, Vol. 1, No. 1, 2015, pp. 15-24, http://dx.doi.org/10.18831/djcse.in/2015011003.

[7] O.Kufandirimbwa, and R.Gotora, Spam Detection using Artificial Neural Networks (Perceptron Learning Rule), Online Journal of Physical and Environmental Science Research, Vol. 1, No. 2, 2012, pp. 22- 29.

[8] A.T.Sabri, A.H.Mohammad, B.Al-Shargabi and M.A.Hamdeh, Developing New Continuous Learning Approach for Spam Detection using Artificial Neural Network (CLA_ANN), European Journal of Scientific Research, Vol. 42, No. 3, 2010, pp. 511-521.

[9] M.Soranamageswari and C.Meena, “Statistical Feature Extraction for Classification of Image Spam using Artificial Neural Networks, Second International Conference on Machine Learning and Computing, Bangalore, 2010, pp. 101-105.

[10] Rekha and Sandeep Negi, A Review on Different Spam Detection Approaches, International Journal of Engineering Trends and Technology, Vol. 11, No. 6, 2014, pp. 315-318

[11] V.G.Gopika and Neetha Alex, A Secure Steganographic Method for Efficient Data Sharing in Public Clouds Journal of Excellence in Computer Science and Engineering, Vol. 1, No. 2, 2015, pp. 11-22, http://dx.doi.org/10.18831/djcse.in/2015021002.

[12] N.T.Mohammad. A Fuzzy Clustering approach to Filter Spam E-mail [A].

Proceedings of World Congress on Engineering, London, 2011.

[13] A.Perkins, The Classification of Search Engine Spam, 2001.

[14] Google Message Security Postini Services. http://www.google.com/postini/email.html.

APPENDIX

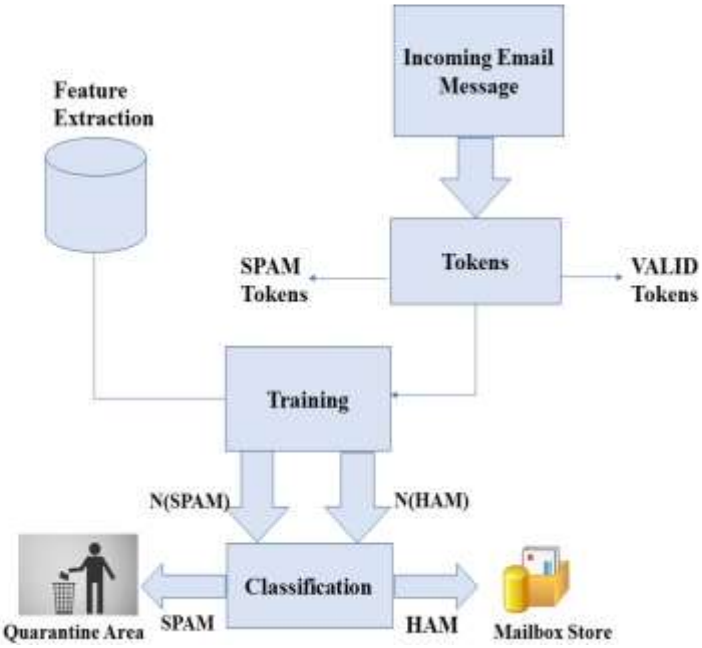


Figure A1. Proposed system

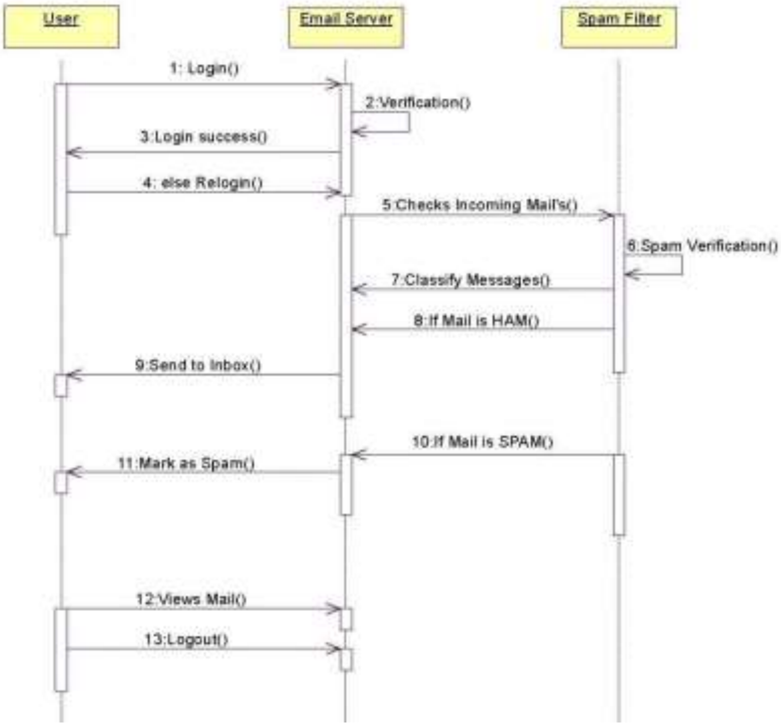


Figure A2.Sequence diagram of proposed system

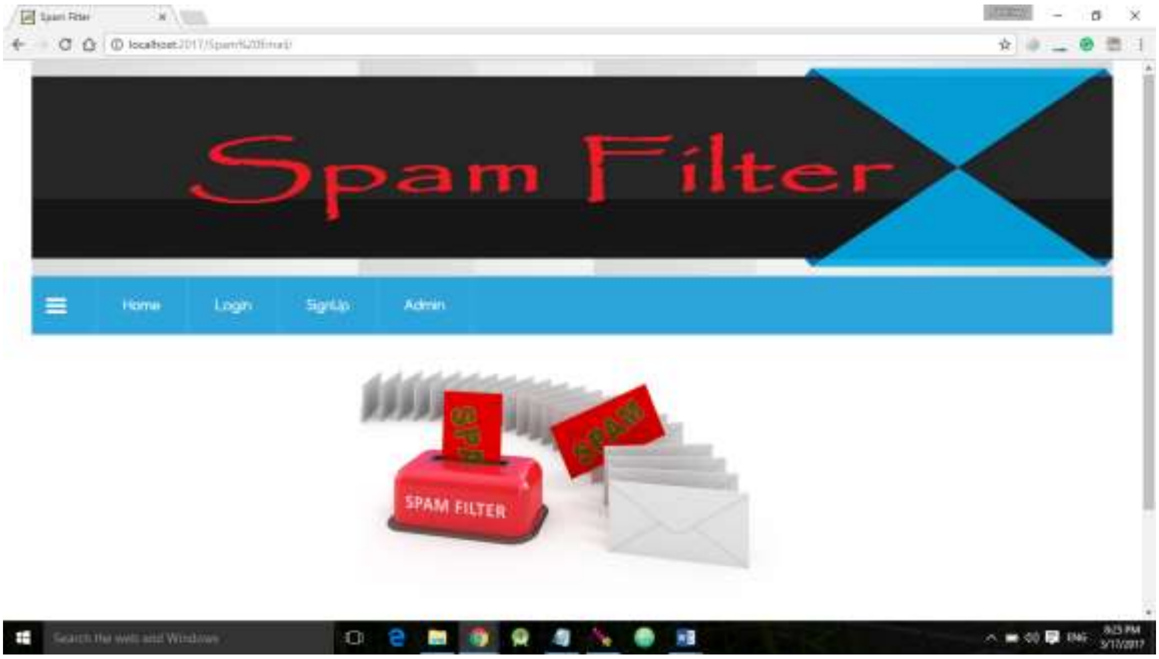
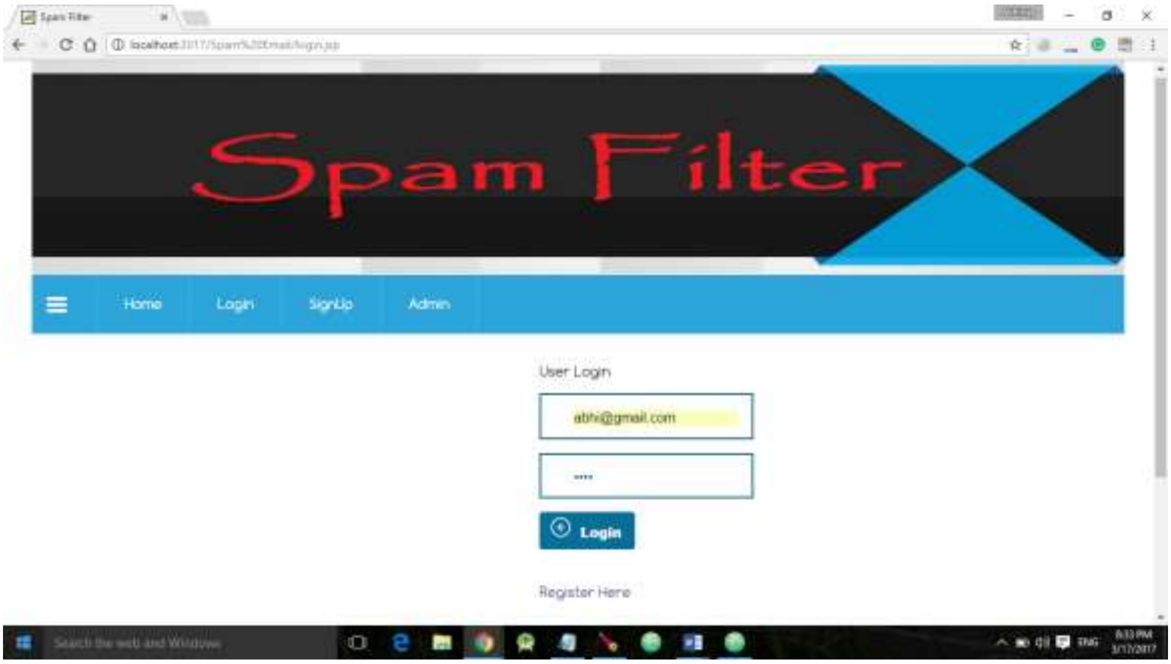


Figure A3.Spam filter home screen

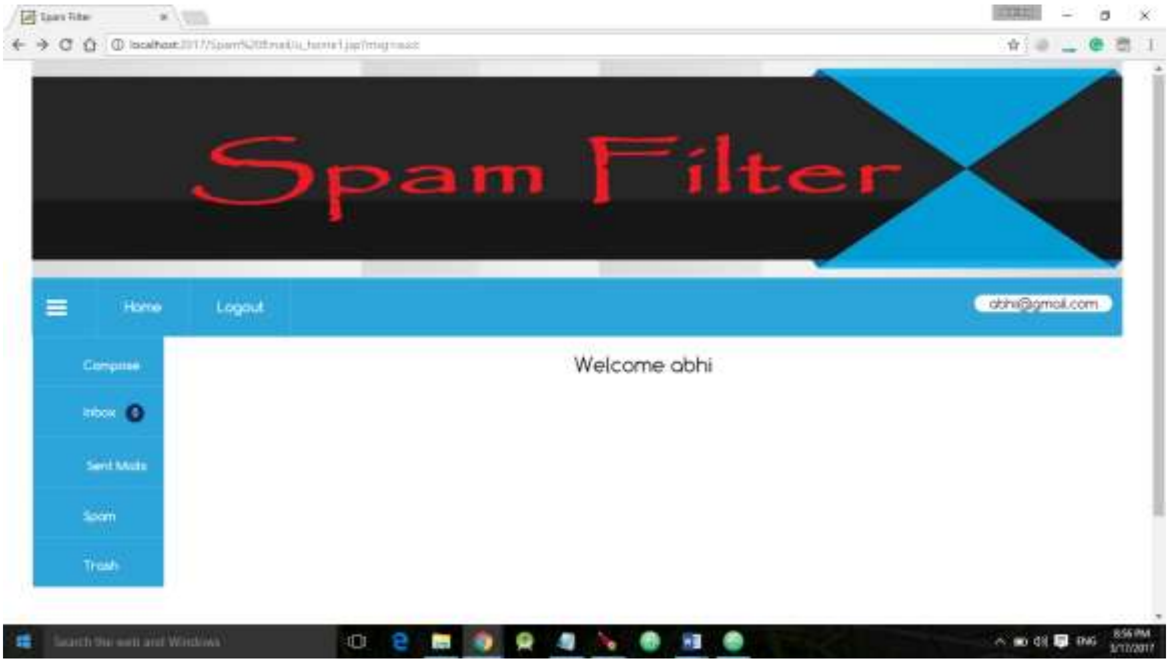


User Login

abhi@gmail.com

****

Figure A4.User login



Welcome abhi

Fig A5.Welcome screen

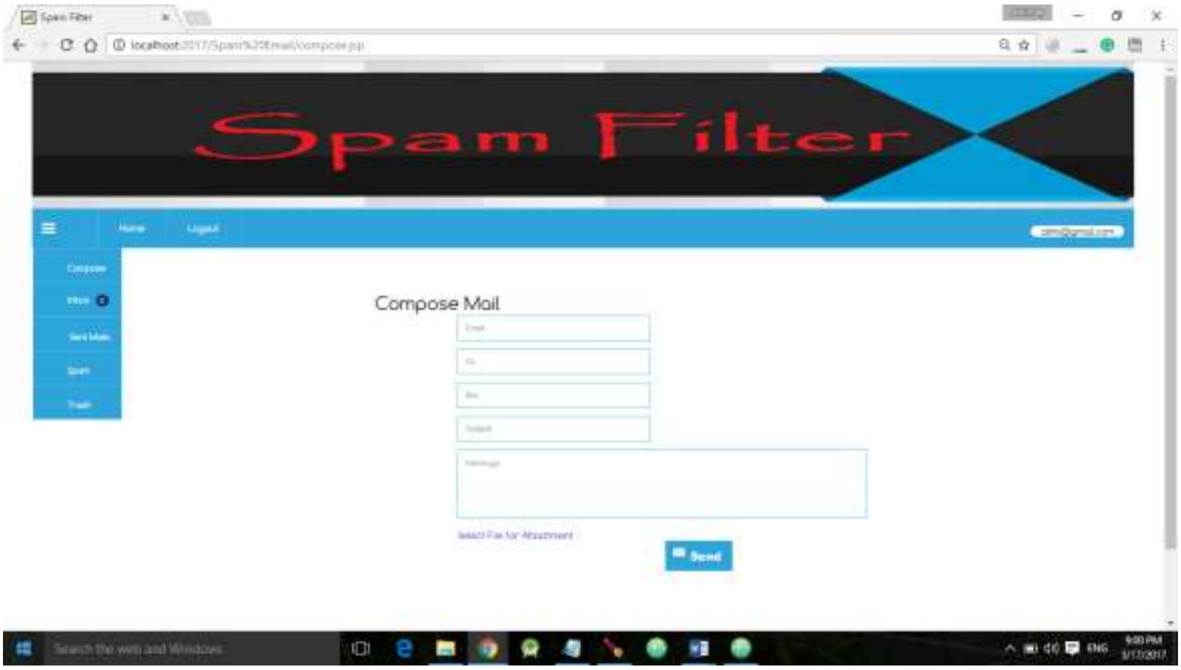


Figure A6.Mail compose

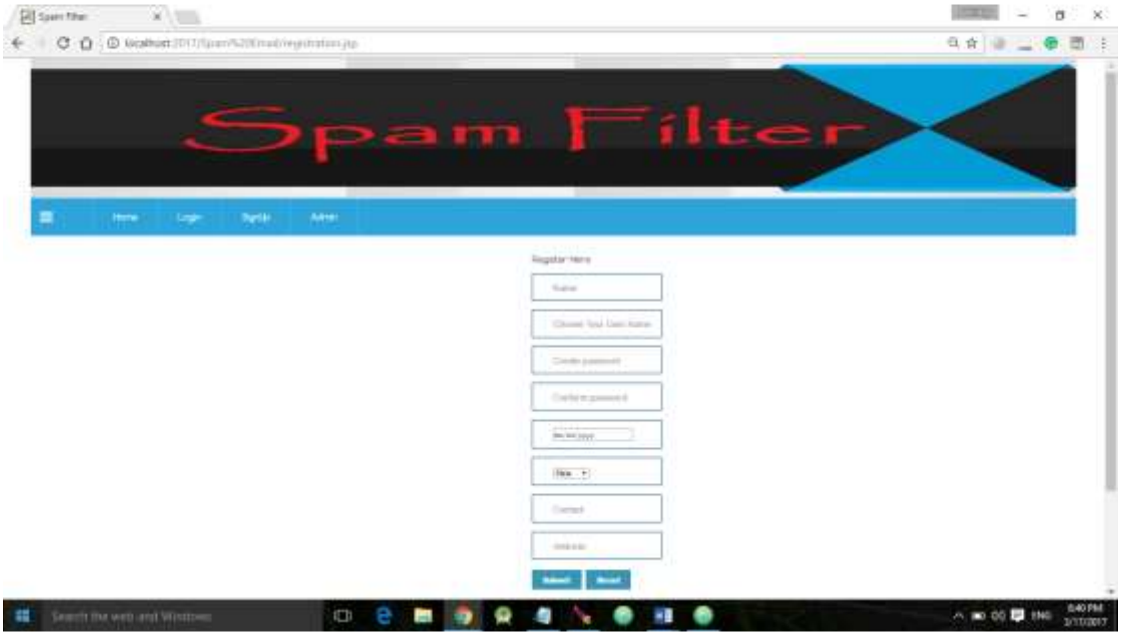


Figure A7.Register login

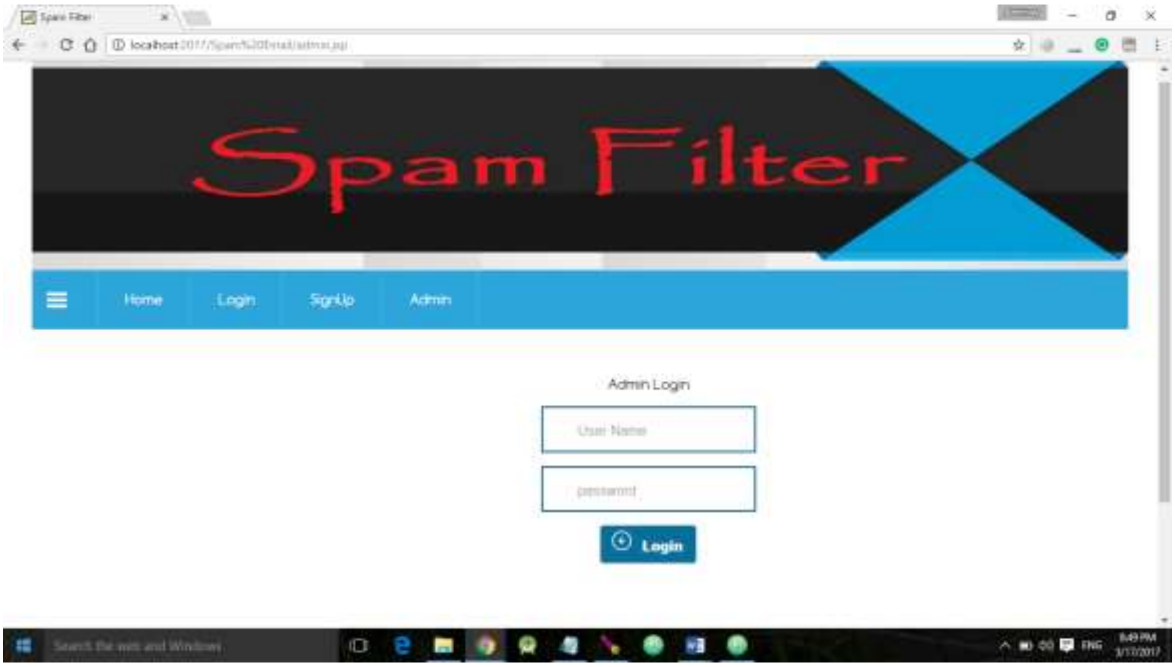


Figure A8.Admin Login

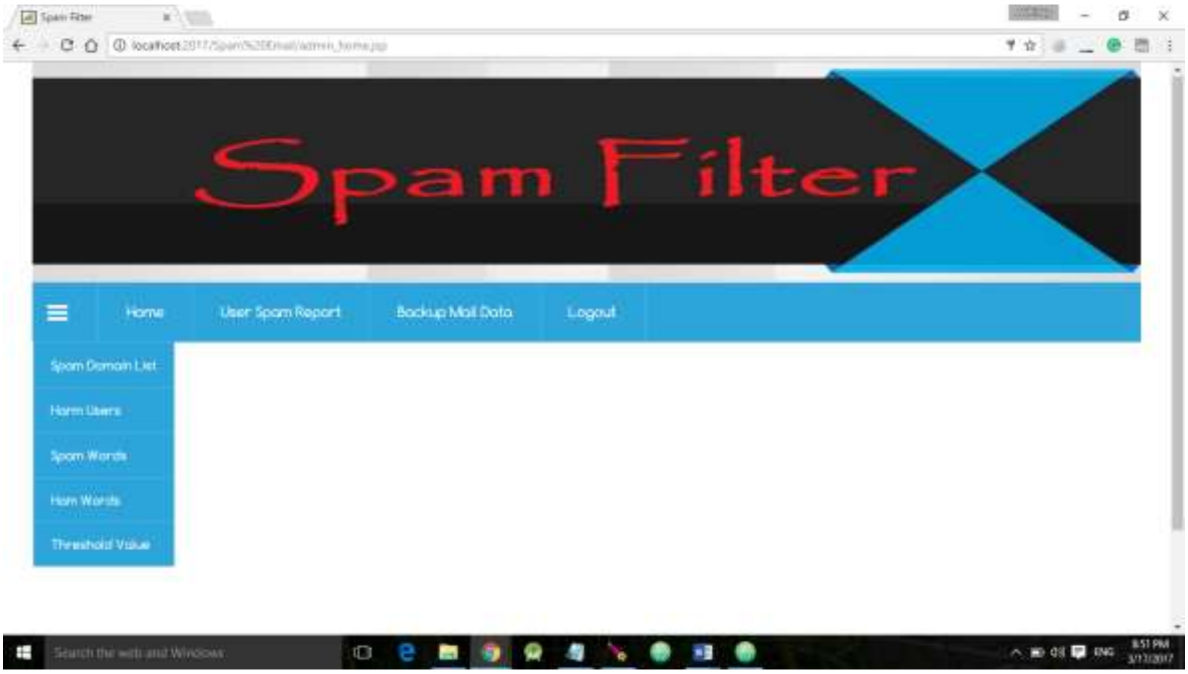
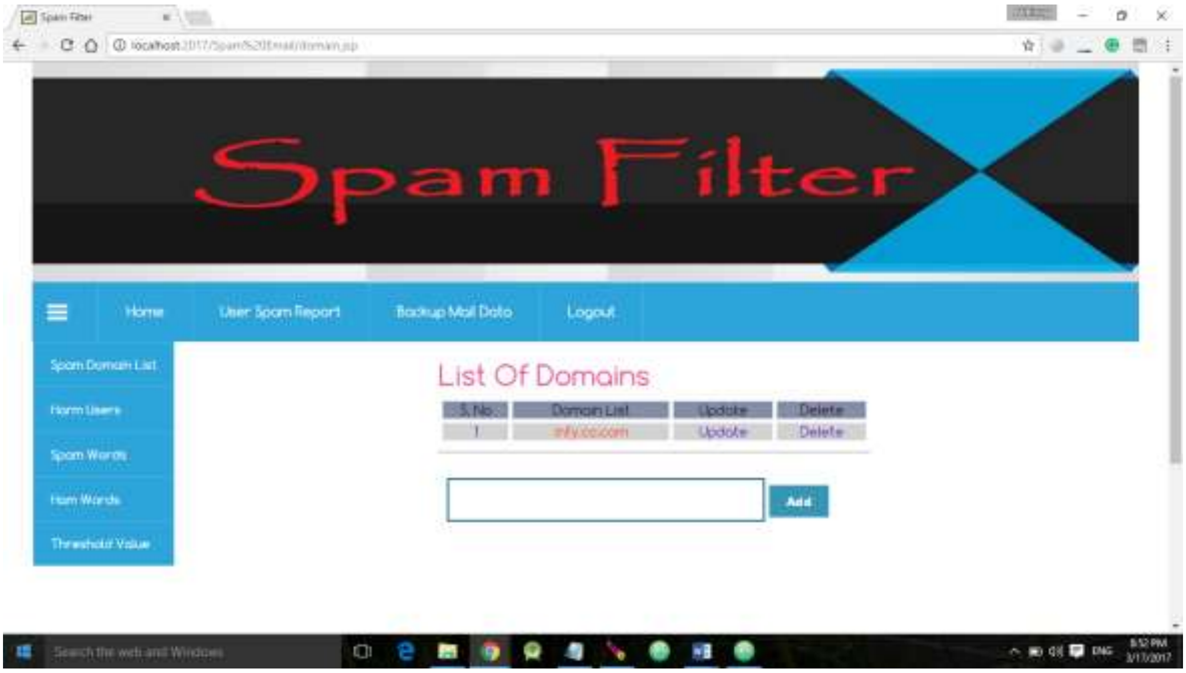


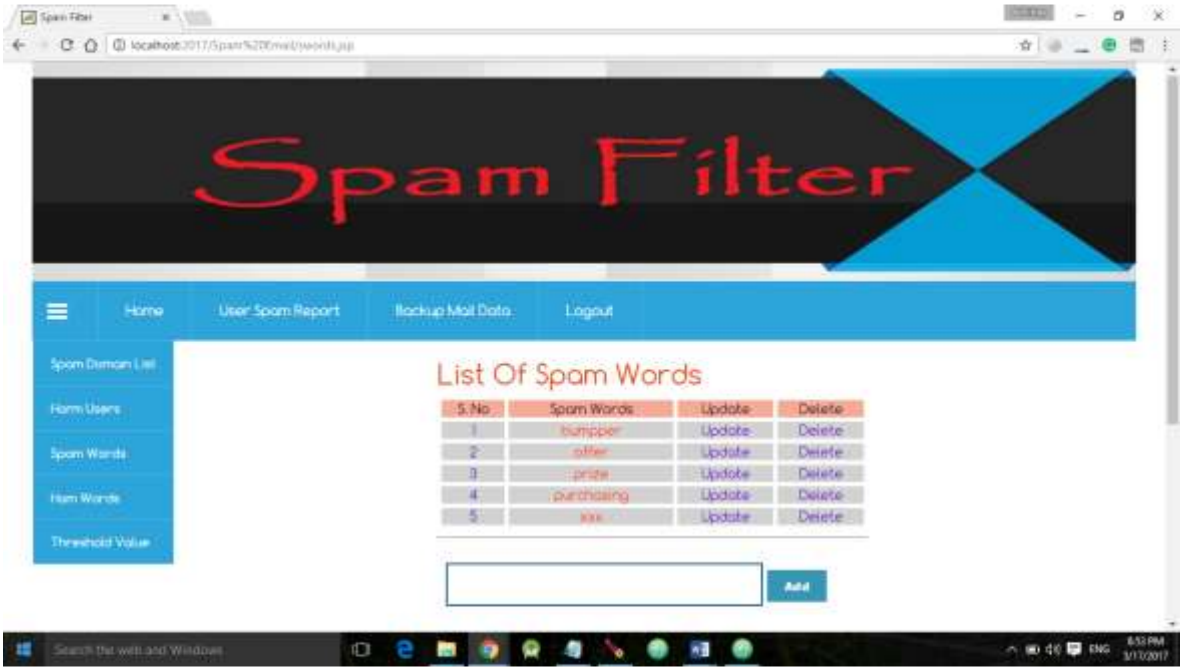
Figure A9.User view



List Of Domains

S. No.	Domain List	Update	Delete
1	infy.co.com	Update	Delete

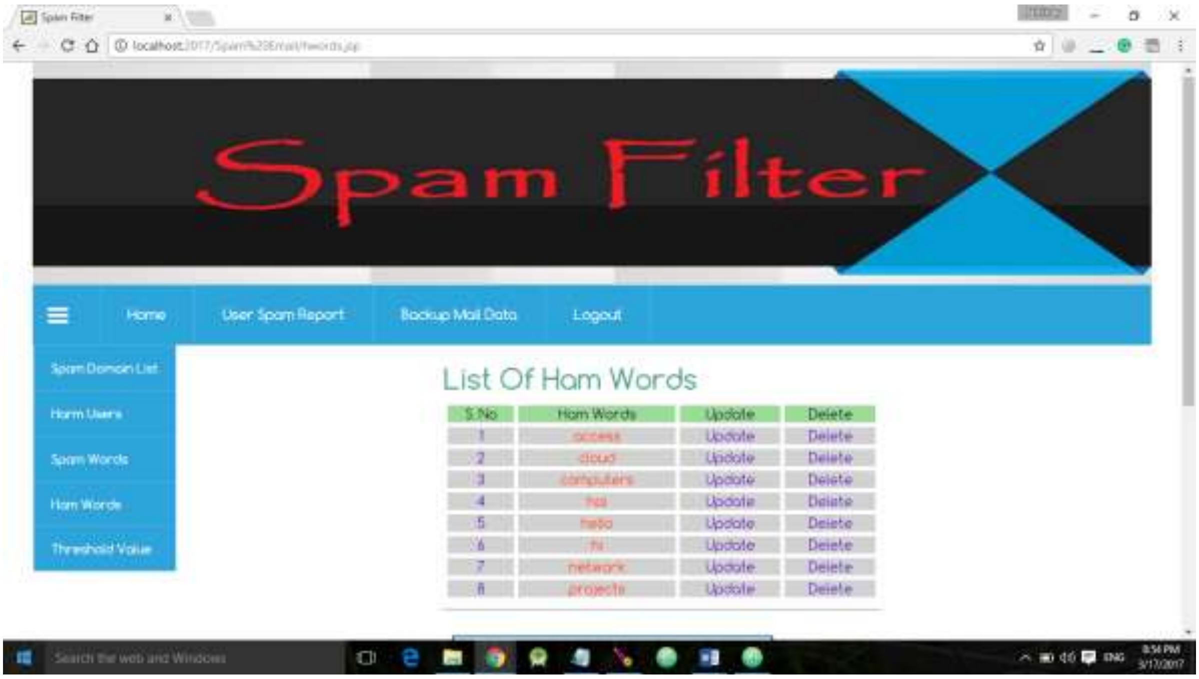
Figure A10. List of domains



List Of Spam Words

S.No	Spam Words	Update	Delete
1	bumpper	Update	Delete
2	offer	Update	Delete
3	prize	Update	Delete
4	purchasing	Update	Delete
5	xxx	Update	Delete

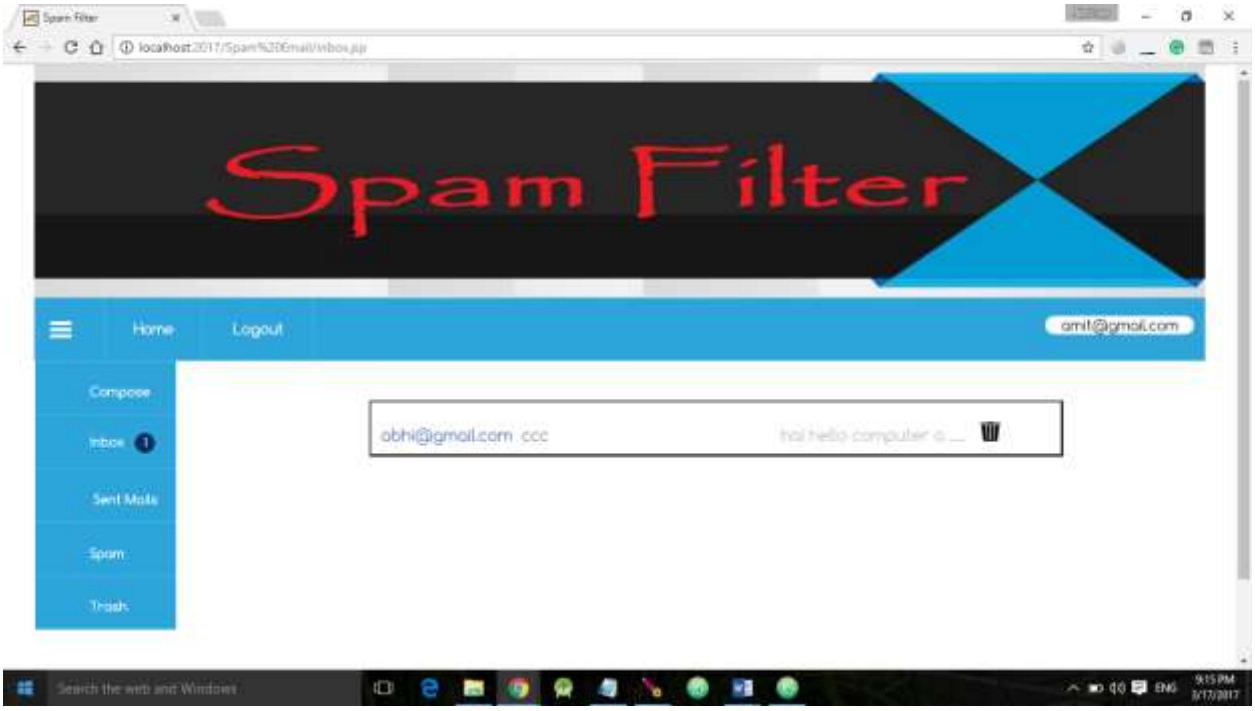
Figure A11.List of spam words



List Of Ham Words

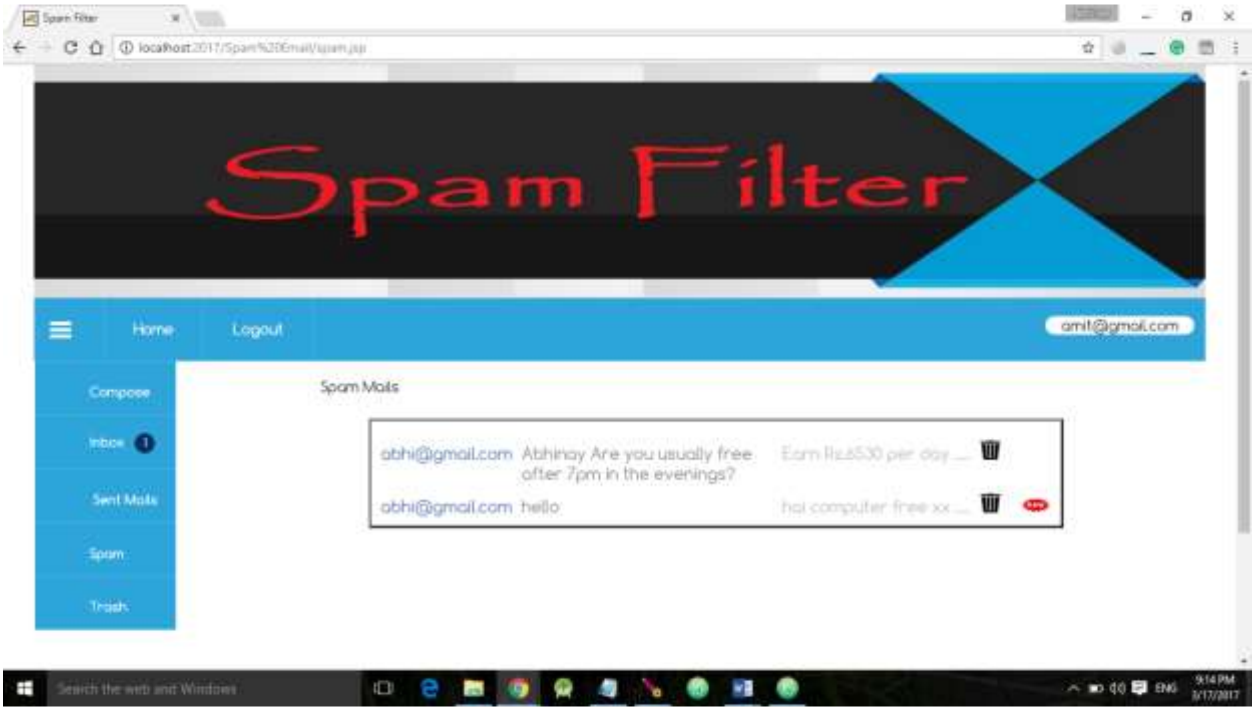
S.No	Ham Words	Update	Delete
1	access	Update	Delete
2	cloud	Update	Delete
3	computers	Update	Delete
4	hai	Update	Delete
5	hello	Update	Delete
6	hi	Update	Delete
7	network	Update	Delete
8	projects	Update	Delete

Figure A12.List of ham words



abhi@gmail.com ccc hai hello computer o ...

Figure A13.Inbox



Spam Mails

abhi@gmail.com Abhinay Are you usually free after 7pm in the evenings? Earn Rs.6530 per day ...

abhi@gmail.com hello hai computer free xx ...

Figure A14.List of spam mails