SPAMKILLER: A
TOOL FOR DETECTING AND FILTERING SPAM

Submitted to the Faculty of
The School of Engineering and Computing Sciences
Texas A&M University-Corpus Christi
Corpus Christi, TX

In Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computer Science

By

Divya Tara Puvvula
Spring 2012

Committee Members

Dr. Mario Garcia
Committee Chairperson

Dr. Ajay Katangur
Committee Member

Dr. David Thomas
Committee Member
ABSTRACT

With the growth of internet email services, spam has gained potential in creating enough trouble to email users by flooding user’s inbox. Efforts have been put in the direction of minimizing spam mails since several years. In light of this need, spam filters have come into existence. Several open source spam filters which are available online have come into existence but spam filters with best features only come at a cost. Basic versions of such filters have drawbacks that could be advantageous to spammers in having their spam mails to pass through these filters unnoticed. Also, commercial email providers have powerful spam filters implemented but these spam filters are not available online to be used by users on other email accounts. The users are therefore forced to pay cost to have spam filters that are effective enough in identifying spam. SpamKiller is an effective open source spam filter proposed that has selected unique features chosen from several online available spam filters.

The proposed spam filter is a web based application wherein the user can login with any of the email user account credentials using which a connection is established with the email server. Several features like white listing, grey listing and blacklisting are implemented to make sure that the fetched mails from the email account are checked and redirected to user inbox or spam folder respectively. User is given utmost capability to customize the spam filter functioning as per user’s preferences. Every incoming mails subject and body is parsed to know about the genuineness of a mail. Through this document, a detailed illustration about the features and functionality of the spam filter implemented are discussed.
# TABLE OF CONTENTS

Abstract ........................................................................................................................................... ii

Table of Contents ......................................................................................................................... iii

List of figures ............................................................................................................................. vi

1. Introduction ............................................................................................................................. 1

   1.1 Email spam ......................................................................................................................... 1

   1.2 Types of spam .................................................................................................................... 1

       1.2.1 Cancel-able use net spam ......................................................................................... 1

       1.2.2 Email spam ................................................................................................................ 1

   1.3 Problems with spam .......................................................................................................... 2

       1.3.1 Unsolicited spam mail come at cost ........................................................................ 2

       1.3.2 Malware spreading/ Phishing ..................................................................................... 2

       1.3.3 Blank spam mails and Forwarding spam mails ..................................................... 2

       1.3.4 Comparison to obtain mailing lists .......................................................................... 3

       1.3.5 Wastage of resources ............................................................................................... 3

       1.3.6 Garbage/ Not legal data ............................................................................................ 3

   1.4 History .............................................................................................................................. 4

       1.4.1 Email spam filter ...................................................................................................... 4

       1.4.2 Types of filters .......................................................................................................... 5

           1.4.2.1 Header filter ..................................................................................................... 5

           1.4.2.2 Language oriented filter ................................................................................ 6

           1.4.2.3 Content based filters ....................................................................................... 6

           1.4.2.4 Permission based spam filter ......................................................................... 7
1.4.3 Open Source Models ................................................................................... 7
  1.4.3.1 SpamFighter email spam filter ............................................................ 8
  1.4.3.2 Drawbacks ......................................................................................... 9
  1.4.3.3 Mail Washer email spam filter .......................................................... 10
  1.4.3.4 Drawbacks ....................................................................................... 11
  1.4.3.5 Spamfence ....................................................................................... 12
  1.4.3.6 Drawbacks ....................................................................................... 14
  1.4.3.7 Spamihilator .................................................................................... 15
  1.4.3.8 Drawbacks ....................................................................................... 17
  1.4.3.9 Spam Assassin .................................................................................. 17
  1.4.3.10 Drawbacks .................................................................................... 19

1.4.4 Rationale .................................................................................................... 19

2. Narrative .......................................................................................................... 21
  2.1 Problem statement ................................................................................... 21
  2.2 Scope ......................................................................................................... 22
  2.3 Functionalities .......................................................................................... 22

3. Proposed system .............................................................................................. 23
  3.1 Open source email spam filter ................................................................. 23
  3.2 Framework ............................................................................................... 23
  3.3 Step by step process of project development ......................................... 24
  3.4 System Requirements ............................................................................. 25
  3.5 System Architecture ............................................................................... 26
  3.6 Flow of execution .................................................................................... 26
3.7 Proposed mechanism.................................................................27

4. Functionality of application........................................................28

  4.1 Modular description .............................................................28

5. Testing and evaluation..............................................................48

6. Conclusion .................................................................................68

Bibliography and References..........................................................69
LIST OF FIGURES

Figure 1: Implementation procedure of eXpurgate technology in Spam Fence spam filter .................................................................................................................................13

Figure 2: Proposed architecture of the SpamKiller representing its modules ..............26

Figure 3: Flowchart indicating flow of execution of SpamKiller........................................27

Figure 4: SpamKiller user interface ‘Settings’ option screen...........................................29

Figure 5: Screen displaying the entries added into the white list ....................................30

Figure 6: Screen showing options to add, edit or delete the entries in white list ............31

Figure 7: List of email ids that are added into the blacklist..............................................32

Figure 8: Operation of adding a new entry to the blacklist is shown .........................33

Figure 9: Words added to the blacklist are displayed......................................................34

Figure 10: Adding new entry to the blacklist operation is displayed .........................35

Figure 11: List of phrases added to Black list of phrases is shown...............................36

Figure 12: Grey list maintenance in the ‘Spam’ option of the user interface................37

Figure 13: Email Ids of first time spam mail senders maintained in the grey list ..........38

Figure 14: Bayesian filter option present in the ‘Spam’ folder .......................................39

Figure 15: Bayesian filtering showing tokens and respective count .............................40

Figure 16: URL’s added to black list of URL’s are displayed .........................................41

Figure 17: Options to edit delete or add new URL’s is shown.......................................42

Figure 18: The inbox of the interface developed is shown.............................................43

Figure 19: The ‘Spam’ folder of the interface developed is shown....................................43

Figure 20: Spam statistics option showing mail statistics ...............................................44

Figure 21: Blacklisted data along with evaluated spam percentage ..............................45
Figure 22: Display of Spam mails with brief subject and sender information ..........45
Figure 23: Display of spam mails when opened.................................................46
Figure 24: The display screen for ‘Settings’ option of SpamKiller............................46
Figure 25: ‘Trash’ screen that contains any deleted mails is shown .......................47
Figure 26: Login screen where the user’s credentials need to be entered ...............49
Figure 27: Adding email addresses to the white list is shown..................................50
Figure 28: Screen showing emails sent from addresses listed in white list in inbox......51
Figure 29: Email address added to the black list is shown .....................................52
Figure 30: Mail received from ids listed in black list shown in spam emails ..........52
Figure 31: Words added to Black list word is shown .............................................53
Figure 32: Email received with the word listed in Black list words shown in spam emails .................................................................54
Figure 33: Phrases added to Black list phrase are shown .......................................55
Figure 34: Mails with the phrases listed in black list phrase are shown in spam emails .................................................................................................................56
Figure 35: New URL added to the blacklist of URL..............................................57
Figure 36: No mail with blacklisted URL is observed in the inbox .......................57
Figure 37: First spam mail from new id is added automatically in grey list ..........58
Figure 38: As first spam mail from new id, mail not added to ‘spam mails’ ..........59
Figure 39: Mails from grey listed sender moved out of inbox after second spam
Mail .....................................................................................................................60
Figure 40: First second mail from new spam sender are shown in ‘spam mails’ .........60
Figure 41: Grey listed mail id added to blacklist on receiving second spam mail ......61
Figure 42: Percentage of spam mails and secured mails shown.................................62

Figure 43: Few mails are available in inbox as per statistics........................................62

Figure 44: Recent blacklist phrase entries with previously stored blacklist data ..........63

Figure 45: Recently added URL with previously stored blacklisted data ......................64

Figure 46: Spam mails identified are shown in ‘Spam’ folder........................................65

Figure 47: Spam mails that are parsed during Bayesian filtering are displayed..............65

Figure 48: Spam mails that are parsed during Bayesian filtering are displayed..............66

Figure 49: Bayesian filter words maintenance with respective count and percentage .....67
1. INTRODUCTION

Electronic mail (email) has become a promising component for Internet users. The most common problem observed in maintaining email inbox is the incoming spam mails. Spam or junk mail is bulk mail which is forcibly sent to many users at once, which otherwise is not accepted by the users. Many copies of the same mail are sent to many users at once.

1.1 Email Spam:

Spam mails flock user email inbox causing difficulty to the user in differentiating between useful mails and spam mails. Spammers are people who spread spam willingly across the internet. Spammers usually obtain email addresses by getting hold of mailing lists from public chat rooms, websites etc. Spammers succeed in finding new techniques to obtain mailing lists from different websites. Unwanted bulk messages can be sent for several purposes. Spam mails usually observed contain information about commercial products, schemes, chain messages or websites that spread malware.

1.2 Types of spam:

1.2.1 Cancelable use net spam: In this type, a spam mail is sent to few users who participate or use newsgroups but never give away their information. A “cancel message” is sent informing not to access usage to the UseNet newsgroups. This message is sent continually till the user inbox is filled with such mails.

1.2.2 Email spam: Email spam is sent directly to email addresses of individuals. These addresses are obtained as mailing lists by spammers from websites, public chat rooms, UseNet postings etc. These mails are usually unwanted mails from anonymous users sent to spread information about products or schemes.
1.3 Problems with spam:

1.3.1 Unsolicited spam mail come at cost: Spam in other words can be termed as unsolicited junk mail. And such mail usually costs a lot to the email providers. Organizations spend a lot of time identifying whether or not a mail is spam before letting it into organization’s email inbox. All this time wasted, costs the company a lot more than the amount spent in initially sending out the spam mail by the spammers. Also, bandwidth consumption for sending and receiving such mails is high. Thus, spam mails barge in as bulk and cost a lot.

1.3.2 Malware spreading/ Phishing: Spam mails create a lot of problems when spread across the internet. Many of the spam mails in circulation carry website links, that on clicking redirect to foreign sites that could be hazardous to the computer. Redirecting users control to phishing sites is most common result of such mails. Confidential information of user is requested through the ‘data fields’ of such sites using which spammers obtain important personal information such as credit card information of the users.

1.3.3 Blank spam mails and Forwarding spam mails: Spam mails can be blank mails without any information in the body as well in the subject of the mail. Sender information is also made unavailable to the end recipient. Blank emails are sent by the spammers which enables them to differentiate between valid email address and invalid address under an email provider. Invalid addresses mails bounce back thus providing spammers with only valid email addresses to further send spam mails. Blank mails sometimes also spread malware which can harm the data in the user’s computer. Trojans in the form of attachments are sent.
Forwarded mails are another problem leading to explosion of spam. In some spam mails, spammers initialize a spam mail and send it to few users, in order to stop receiving similar spam mails further, the user is forced to forward the mail to some others in the mailing list. Hence, even if half the users forward the mail, the amount of spam created is immense and would require lot of cost to be removed off the internet.

1.3.4 Comparison to obtain mailing lists: Personal email addresses could be out open in the wrong hands of spammers using spam mails. Spammers sometimes obtain access to database containing the personal information of users. If they obtain another database containing some other information like email addresses, spammers compare the databases to extract subtle information of the users and then flood the inbox of the user with unwanted mail.

1.3.5 Wastage of resources: Spammers usually gain access to genuine intermediate resources that are known to be benign by users. Spam mails are forwarded with these intermediate resources as the ‘sender’. These intermediate resources are initially flocked with spam mails and are then forwarded further to individual users. As a result, these intermediate resources are defamed for spreading spam and are held responsible. Sometimes spammers create new administrator accounts and send out spam mails after which the account is deleted or is left to the benign providers who would be named responsible for spreading the spam mails.

1.3.6 Garbage/ Not legal data: Most of the spam mails prevalent are junk mail consisting of nothing that is worthy to the user. Spam mails usually contain information about schemes and products that are not of much use to the individuals. Fraudulent schemes, solutions for situations,
free advice, links to phishing websites etc. are sent through spam mails which mostly tend to be only garbage material. Content that is considered illegal on Internet is also spread via spam mails. In certain countries, laws are implemented against display or spread of certain content. Spammers, against those laws, try to spread out content that is considered illegal through spam mails and create trouble for the end recipients.

1.4 History:

1.4.1 Email spam filter:

A solution proposed way back when Internet came into existence was to implement spam filters [16] to avoid spam from filling email inbox to the brim. Improving existing spam filters has been an ongoing war against spam mails ever since email services gained importance. The main purpose of spam filters is to process the incoming mails and organize in accordance with user preferences. Without the implementation of spam filters or blockers, individuals are forced to take out time to check if the flood of incoming mail is spam or not and then delete them manually which is a tedious task.

Usually, the preferences of user are maintained and hence, on every incoming mail that is accepted as input by the spam filter, these preferences are automatically taken into consideration when the mail is processed. If the incoming mail is identified as spam, the filters take necessary action either to delete it or to move it some other folder. Spam filters can be installed on individual workstations or can be directly implemented on the email clients to have the incoming mail checked to avoid spread of malware or spam into inboxes. For every incoming mail, spam filters take help of parameters like sending network IP address, pattern of words common in
spam mails and so on, to recognize with accuracy whether a mail is benign and also whether or not the recipient is the destined correct recipient of the mail.

Email spam filters are implemented by default on every email client. Even then, many users still prefer to run spam filters individually on workstation to avoid being flooded by spam in inbox of personal email accounts. Email spam filters with high degree of accuracy and best features usually come only at a cost. Open source email spam filters are also available for free of cost but lack to have the best features. Many filters that are open even if installed on user’s workstation do not effectively mitigate the flow of incoming spam.

1.4.2 Types of filters:

Development of spam filters should be such that it is customizable as per user’s criteria. Flexibility to work in accordance to individual’s whims would be the best feature to be implemented on a spam filter. A reliable email spam filter that would work no matter what type of spam mail come in should be the main motto in developing effective spam filters. Several techniques have been implemented in spam filters to identify spam. Some such filters are:

1.4.2.1 Header filter: Checking header of the incoming mail to identify if the mail is “spam or not” has been the first type of spam filters implemented. Headers usually possess information like address of sender or sending network, recipient address etc. When a mail arrives, the spam filter gets hold of the header information of the mail and checks whether or not it is genuine enough to forward to the user inbox otherwise trashes it. Information of servers that are involved in sending the mail can also be retrieved from header information.
Hence, for every incoming mail, the header information is obtained and is checked against all odds to identify if in case a spam mail is on spree. Spammers use different techniques to mislead spam filters to let spam mails in, but header filters would be developed to be smart enough to identify any forged headers. Sender information is checked to see if it is from a legitimate accessible sending network or from a spammer intending to spread spam mails. But, implementation of this mechanism only, on a spam filter wouldn’t be enough when benign known sending networks try to spread spam.

1.4.2.2 Language oriented filter: Language oriented spam filters are another type of filters implemented that do not give the best of results in identifying and deleting spam. Hence, helps only in avoiding some spam mails from entering the system. Users set up mailbox preferences in accordance to their convenience and comfort in using it. Users across the world prefer different languages as the key language in which the mails and the email interface is made visible to them. Incoming emails are also usually sent in the same language as the preferred language of the user mailbox settings.

Spam filters developed with language filtering as the main mechanism remove out any incoming mail that is in any other language besides user’s mailbox language preference. Spam mails usually are not sent according to individuals preference instead are sent in bulk to many users at once. Only spam mails sent in other languages are filtered out and hence could be of little use when implemented along with features of header filter and other filter features too.

1.4.2.3 Content based filters: As the name suggests, content based filters judge an email based on the composition of the email. Content based filters are the most common type of filters that are on high these days. For every incoming email, the contents of the email are
scrutinized and then a decision is made whether or not the mail qualifies to be a spam mail or not and necessary actions are taken.

Spam mails usually exhibit common patterns. Hence, a note of such similar patterns common in majority of the spam mails are maintained to help in further judging incoming mails to inbox. Usage of certain words is most frequently observed in many of the spam mails in circulation. When a mail comes into a content based spam filter, the data in the mail is scrutinized and is evaluated against the patterns of spam mails maintained. A decision is then made after performing all possibilities of checking a mail and then the respective action is performed based on the decision. This type of filters cannot be accurate sometimes as long emails or letter based emails are wrongly identified as spam and are quarantined sometimes.

1.4.2.4 Permission based spam filter: Permissions to allow or stop mails with certain qualities can be personalized and customized by the user. Permissions and privileges are set up by the user to allow or deny entry of certain mails into the inbox thus mitigating the entry of spam mails. These filters allow and permit mails only that are considered authorized by the filter. Authorization can be given to the sending network, header information, presence of certain words etc. All such criteria can be customized by the user. For every incoming mail, such authorization criteria are all taken into consideration for making a decision about a mail and respective action to either trash the email or forward it to the inbox of user as genuine mail is performed.

1.4.3 Open Source Models:

There has been a rapid growth of methodologies that can be applied to mitigate the unpleasantness caused by the spread of spam. Several tools have come into existence and are
made available to the users. Open source models are those that are made available online to the users either at a cost or for free. Some such models are discussed:

**1.4.3.1 SpamFighter email spam filter:**

Spamfighter [20] is named as one of the best email spam filter in use over the internet and works in association with Microsoft. It has successful implementation in identifying spam and avoiding spam from cluttering up the inbox. A separate installation procedure is used to get this tool implemented on user workstation.

- The most advantageous feature of Spam Fighter is the ability to directly import the contacts list of the user email inbox to the maintained white lists. Contacts imported into the white list are considered genuine and any further mails from the imported contact list are directly forwarded to the user’s inbox.

- Community based filtering is employed in identifying incoming mails. Users usually maintain different groups based on social closeness like friends, family etc. Information about user’s social closeness is maintained by the spam filter. White lists are formed depending on the social closeness of the user with the existing groups of contacts.

  Groups like ‘family’ gain the high importance and hence are added to the white list. ‘Close friends’ could be another group that could also be added without a doubt into the white lists. Lists like ‘Acquaintances’ are given next priority and so on. Whenever a mail comes in, the filter identifies the entry in the white list, if present in the white list, the entry is searched in different groups and based on the closeness of the group to the user, a decision about the mail is taken. If the mail belongs to close groups like friends or family, the mail without a thought is forwarded to mail inbox. If the mail is from not so close groups, the mail is further checked for being spam and decision is taken.
● User accounts need to be created in order to access the features of the SpamFighter filter. Another best feature of this tool is the identification of mails with phishing sites. Spammers usually send out mails with hyperlinks to sites that are phished or induce Trojans into user systems. The incoming mail is parsed to identify presence of any hyperlinks. The words used in these hyperlinks are checked for the presence of words that are most frequently used in spam mails. If any such words are found, probability of the link being forged is high and the mail is identified as spam and is quarantined.

● Many email accounts could be managed by this tool. As it is a standalone application, several email accounts can be handled using the same tool on an individual’s workstation. If a mail is accepted once from a user and is considered genuine, then any mails from the same sender any time further are considered benign and are forwarded directly to the inbox. This feature sometimes cause problem as spammers could make use of this loophole and send spam mails using white list contacts as intermediates. This filter hence at a cost provides unique features to mitigate spam from entering personal workstation.

1.4.3.2 Drawbacks:

The main disadvantage of this filter is its incapability to adapt any new techniques used in spreading spam. Spammers are exploring new techniques to make sure that the spam mails pass through these advanced filters unnoticed and as genuine mails. Many novel methodologies are used to spread spam these days which spread spam despite having powerful spam filters implemented by email providers. No mechanism has been added into this filter that could be trained or get adapted to the changes in spam spread techniques.

Junk mail cannot be stopped from entering the inbox as the contents in the mail are not parsed, instead only the genuineness of the sender is evaluated by observing the social closeness
of the sender with the user. Also, accepting mails from a sender considered as genuine once is a drawback, as it could be used to the best of spammer’s advantage.

1.4.3.3 Mail Washer email spam filter:

Mail washer [8] [21] [22] is one of the most visually pleasing email spam filter in use. It has an interface that could be very easily understood. Mail washer works well with almost all the famous email providers like Gmail, yahoo, AOL etc. Features of the filter are:

- As mentioned, visual interface of this mail filter is very pleasing and straight. All the options available are organized such that it is very easily accessible by the users. All the mails in the email accounts of the user are drawn into a common interface. A clear preview of the subject of the mail, sender of the mail is displayed that helps the user to decide whether to view the mail in detail or quit. Any spam mails that are identified are marked in a different color indicating the user of the detection and it is for the user to decide further about the mail.

- The filter provides absolute control about how the filter has to work to the user. Customization of the filter can be done completely by the user. Lists of senders from whom the mails can be directly accepted can be maintained by the user. Filter with certain words in the subject could be filtered out as spam. This feature enables the removal of junk mails that are repeatedly sent to user inbox. Forwarding mails to some other mail account when a mail is considered genuine or removal of mails when identified as spam can be customized completely by the user. Information about ‘friends’ can be maintained in the contacts list. Every time a mail is received from these groups, mails are considered genuine and are directly forwarded to the inbox.
• Deletion of mails when identified as spam is another such option that could be of help and as well cause problems sometimes. Mails that have the highest probability of being spam are usually automatically deleted by the filter. This option can be customized by the user. If this option is not chosen, spam mails are marked in other color leaving the user to make decision about the action to be taken further on the email.

• One of the best features available for the filter is the Bayesian learning mechanism implemented in the filter. As the methods implemented by spammers in spreading spam change, the more difficult it becomes to identify the spam. But by implementing Bayesian learning, the filter can be trained about the new trends using which spam is sent. The filter is trained with spam mails as input and hence gets versed with the new spam mails. For every incoming mail after the training, the filter applies these techniques in identifying whether or not the mail is spam. Hence, the more the filter is trained, the more is the accuracy in identifying the mails that are spam, and less is the probability that spam mails can make it to the user inbox around the filter.

1.4.3.4 Drawbacks:

There are certain drawbacks that could help spammers in getting spam mails into circulation. White list maintaining records of senders or sending networks that could be considered benign besides the contacts in the friends list is not observed. Mails with forged headers, attachments or images are not identified as spam by these filters.

Texts converted to images are sent as spam these days. Hence, spam mails with such images can go in unidentified through the filter which could decrease the accuracy of the filter. Sender information can be forged by the spammers; the filter would be incapable in identifying
such forged headers. If the sender information is stated as genuine in the list of acceptable senders, spam mails with forged headers can get into the mailbox easily.

1.4.3.5 Spamfence:

Spam fence [6] [9] [11] is one of its kind as it implements the new eXpurgate technology. This filter is renowned for its ability to stop mails that could spread viruses or any other malware. Mails that are not spam are never mistaken to be spam by this filter. Some of the important features of the spam fence filter are:

- This filter has the lowest rate of false positive occurrence. A mail that is benign is never identified as spam thus being helpful in not letting important mails to be lost while filtering spam. Lengthy mails usually tend to have the highest probability of being recognized as spam. In this methodology, lengthy mails are not initially assumed as spam instead header information is checked to know the legitimateness of the incoming mail.

- eXpurgate is the guideline technology used by this mail filter. The implementation of eXpurgate gives a clear idea of the working of spamfence filter. eXpurgate is implemented in three main steps. Working of eXpurgate is as shown in Figure 1:
  1. All the users intending to use this filter need to register as a new user for eXpurgate account. All the mails of any email account (user account 1) associated with this filter are initially forwarded to this account. That is, for the incoming mails to be filtered, the mails are forwarded to the eXpurgate account.
  2. In the second step of implementation, the mails are checked. All the header information of the incoming mail is taken into consideration and the genuineness of the mail is checked. An extra additional header is added to the
incoming mail. The extra header appended to each mail indicates the ‘result’ of the filter check done over that mail. A decision about the mail like ‘spam’ or ‘clear’ or ‘suspicious’ is made and is appended as the header to the mail. This mail is then forwarded to another email account that is the prioritized account (user account 2). This user account is the one that is being saved from the entry of spam mails and other malware junk.

3. The mail with the new header added is then forwarded to other email account which is deprived of spam. Here, the additional header is parsed to know the status of the mail. If the mail header indicates that the mail is ‘clear’, that is in other words, if it is considered as benign, then it can be directly sent to the inbox. The filter at the second email filters out any mail that has the header information as ‘spam’. And for mail that is marked as ‘suspicious’, a second check is done to confirm whether or not the mail is spam. Thus, this decreases the scope of false positive occurrence.

Figure 1: Implementation procedure of eXpurgate technology in Spam Fence spam filter [6]
• Bulk email can also be checked at once using this filter. Bulk of mails are checked once by the eXpurgate filter and are forwarded to the mail account where a second round of checking is performed. Also, a common header can be added to a bulk of mails and can be forwarded. All the mails under a common header obtained are checked and a common action is performed on all the mails. This feature thus reduces the effort on the filter as it avoids the functioning of the filter on each and every email as it can be worked on as bulk.

• Customization is another feature that can be used to its best by the user. Complete working of the filter can be as per the user preferences. Settings and options are provided such that the user decides the action that need to be performed on each kind of header information. Options to direct mails that are marked as spam to trash automatically are also made available to reduce effort on user to delete them manually. Mails with different headers can be forwarded to different accounts on user’s request. Hence, capability to customize the filter totally is provided to the users.

1.4.3.6 Drawbacks:

Adaption of the filter to the new spam techniques is not observed in this filter. The filter cannot be trained with new spam samples so that it can get adapted to the trend of spam mails in circulation. Spammers engage in new techniques to get going with spam mails even though spam filters are implemented by email providers. Image spam is also not taken into consideration while parsing the mails. Images constituting of lot of text is not parsed and hence such spam is free to enter the system.

One of the major drawbacks is the need of two accounts in order to forward genuine mails. Once the basic checking is done at the eXpurgate account, the mail with the extra header
is forwarded to another account which needs to be mentioned to have spam less mails. This could be a problem as the user is forced to maintain two accounts of which only one account can be made spam free. Also, mails are only checked and not analyzed based on the sender information. Only mails are parsed and checked for any spam content. Hence, mails from spammers cannot be identified using the sender information.

1.4.3.7 Spamihilator:

Spamihilator [23] is one of the several other open source filters that are available online. It has several features implemented that could bring positive results in identifying spam. This filter is implemented between the email client and the internet. Some of the features of this filter are:

- This filter can learn and get adapted. It can be termed as a learning filter. This filter can be trained as the spam being spread using different methodologies. The varied the range of the spam used to train the filter, the more efficient the filter becomes in identifying the upcoming spam. Bayesian learning methodology is implemented using which the filter is always up to date with the current spam in circulation. Probabilities of the mail being spam are evaluated and then a decision about the mail is taken. A threshold value is fixed and any mail with a probability value more than the threshold is considered as spam and if the probability value is less, it is considered genuine.

- Content based filtering is encouraged in this filter. Certain patterns of words most commonly appear in spam mails. Some such words are, “You won”, “Viagra”, “Win money”, “Child pornography” etc. All such words have probability of appearing in spam mails more often. All such probabilities are made known to the filter. A database containing of words that are frequently observed in spam are maintained. This database can be extended as the possibility of words grows.
Whenever a mail comes in, the words in the mail are parsed and are compared against the words in the database. Probability of the mail being spam is evaluated by evaluating the probabilities of the words in the mail. A final probability value is obtained and then decision whether a mail is spam or not is made.

- Separate lists of users with different privileges can be maintained in this filter. In other words, list of addresses that are considered genuine are maintained in the white lists. The database can be extensively expanded to support the need of the individual user. Any incoming mails with header information matching the addresses in white list are considered benign and are directed to the user inbox.

- Also, another list called black list is maintained containing records or information of senders identified as troublemakers. Blacklist maintenance is one of the best features implemented as it could stop or block mails from users whose mails are not intended to be seen by the individual. Users are given complete ability to block mails from users that are assumed to be spreading spam mails.

Many sending networks are known to send spam mails and hence such networks addresses are identified to be placed in the blacklist. Implementation of blacklist could be of great advantage in mitigating the spread of spam. Also, personal choices of users can be satisfied when mails need to be stopped from unknown senders. Customization of the records maintained in blacklist is provided to the user. Blacklist database can be extended as per user requirements.

- For every incoming mail, check against white lists and blacklists is done and a decision about the mail is made. All the mails with senders identified in blacklists are directed to the ‘trash’. Ability to restore the mails that have been identified as spam or that have been
trashed is given to the user. Options are provided where in user can decide the action that needs to be taken for mails identified as spam. Redirecting option is made available in case users want to forward genuine mails to other mail accounts.

1.4.3.8 Drawbacks:

Despite having many advantages still there are drawbacks that could be used by spammers in spreading spam. The most common type of spam that circulates in internet is image spam. Large amounts of text in spam mails are changed into an image and are sent to several groups of users. Such mails need to be filtered out by spam filters before delivering the mail to the inbox. This filter does not have any mechanism implemented to identify image spam. Hence, spam mails with images are not identified which could help spammers to send out more spam mails.

Mails are checked for spam using sender information in the white lists and blacklists. But spammers use techniques where mails are sent using forged headers. No mechanism is implemented to identify forged headers from spammers. Hence, spammers can use genuine intermediate email providers to spread spam mails.

1.4.3.9 Spam Assassin:

Spam assassin [12] is another email spam filter available online that works in association with Adobe. It has got a very easy and flexible configuration setup. All the configurations are available in plain text and hence any modifications can be made easily as per the requirement. Features of Spam assassin are:

- Bayesian filtering is implemented in the filter for identifying spam. Training the filter with all types of new spam would help in adapting to the new techniques used while
sending spam mails. A separate option to train the filter is made available. The more the filter is trained; the best the filter gets versed in recognizing spam mails. Learning and getting adapted to the new spam spreading techniques makes the filter efficient and accurate in identifying spam mails from all the incoming mails.

- Blacklists comprise sender’s information whose mails are to be blocked from entering the mailbox. Blacklist addresses are maintained in the databases that are made available online. Senders whose mails usually tend to be spam are maintained as records and are added to the blacklist. Database maintaining blacklists can be made extensive to remove most of the spam. For every incoming mail, initial check is done against the blacklists and an immediate decision is made about directing the mail. Spamhaus is an online website that maintains lists of network providers that need to be considered genuine as well as networks that need to be blacklisted.

- Content matching is also implemented in this filter. Several patterns of spam mails are already known from observation of spam mail. Mails can be identified as spam based on the content in the mail. Words that are most common in spam mails are maintained in the database with probabilities of occurrence. For every incoming mail, words in the mail are considered and are matched against the database to obtain probability of being spam. The overall probability is used to make a decision about the mail.

- Regular expressions are used to check the body and header fields of a message for higher accuracy. A final score after evaluations is assigned to the mail to help in deciding the respective action on the mail. High probability of the score is considered as spam and is either marked as spam or is deleted. An internal threshold is decided; any mail with probability greater than the decided threshold is considered as spam and is quarantined.
1.4.3.10 Drawbacks:

Adaption to the changes of spam techniques is not observed. This feature is a must in spam filters as spammers change techniques very often to outsmart the filters in email providers. Image spam is another important issue that is on rise these days over the internet. Images containing spam text are sent out by the spammers to flood inboxes of innocent users who otherwise would never want to have such mails in the inbox.

With only the maintenance of blacklist and content based filtering, the filter would not be effective enough in filtering out all the spam mails sent to a user. Spam mails could still make their way through these filters thus not being fault-tolerant sometimes when needed.

1.4.4 Rationale:

With the rise of internet email services, spam has been on a great high since then. Spam mails which are not wanted have always been forced into user accounts. Commercial spam mails as well as noncommercial spam mails about invalid schemes or products are spread across the internet. Immediate solution for the problem of spam was made available in the form of the email spam filters. These have become a must component in all the email services these days. Many different techniques have been implemented in these filters to make sure that spam mails are not delivered to the user’s inbox. Even though many mail filters have come into existence, spammers are still able to come up with new ways of sending spam mails. Spammers make sure that spam mails go unidentified through the filters straight into user’s inbox. Today, even with the best spam filters implemented by commercial email providers, users face the problem with unwanted flood of mails.

Commercial email providers use best of the spam filters which come at a price. Individual users with different email services are forced to use spam filters available online.
These filters are usually termed as open source online available email spam filters. Email spam filters that could be used by individuals are available both for free as well at a cost. Most of the currently available spam filters have the best features available only at a price, and the features provided for free versions are very basic. These features would not be reliable enough in filtering out the spam mails. Email spam filters should be evolving as per the trend to outsmart the techniques implemented by spammers in sending spam mails.

Open source spam filters can be used by individuals for email accounts but, to get rid of maximum of the flood spam mails, a costly spam filter must be used. Hence, there is a need of an email spam filter that could be effective as well be available online at free of cost. Hence, SpamKiller helps by providing best features of some spam filters as ‘all in one’ spam filter that could be effective without a cost.

The results show the effectiveness of SpamKiller in handling spam mails. This project could be extended further by appending more features that could trace out spam mails with even small traits of being a spam mail.
2. NARRATIVE

2.1 Problem statement

Spam filters have been evolving ever since they were first implemented on email servers. Initially, only the sender’s information was checked to know whether or not a mail is genuine. As spammers enhanced the techniques used in spam mails, the spam filters were enhanced too to support the needs of the users. Also, adaption of the spam filters to changes in different spam mails is not observed in many of the new generation spam filters. Hence, many spam filters have come into existence with new techniques. Almost all the commercial email providers these days use powerful spam filters which are not available online. Other powerful spam filters have been implemented with several unique features in detecting spam mails. Not all the filters have the best features in identifying spam. Several such filters with unique features only come at a higher price.

Spammers use smart techniques in sending out spam mails and hence spam filters should be capable enough to identify such smart spam mails. Not all spam filters available online would have greatest accuracy and efficiency. Several drawbacks exist for each of these filters. Online available filters with best of the features are not available for free. Adapting to evolve has become a must for all the spam filters in implementation. Hence, there is a need for a spam filter with unique features that could be made available for free online. Therefore, SpamKiller is developed that has implementation of best features selected from different spam filters. This filter would thus solve the need to pay amount in order to have a personalized spam filter with effective features.
2.2 Scope

The main objective of the project is to develop an email spam filter with different features selected from several spam filters that could be made available online at no cost to individual users. Email spam filter developed should be such that for every incoming mail into a user account, a check has to perform whether a mail is spam or not. If the mail is identified as spam, then the mail is sent to a separate folder, of not the mail is to be forwarded to the user inbox. The filter has to be effective enough to decrement the entry of spam mails into the inbox. Identification of mails with image spam or unauthorized sender or hyperlinks to sites that spread Trojans etc., has to be achieved. The performance of the filter is analyzed by considering the accuracy by which the filter identifies the spam mails. Fault tolerant, reliable behavior is needed for every incoming mail. The outcome of the project is the spam filter that can filter out majority of the incoming spam mail with no cost of usage. The project is tested with different mails, of which some are spam mails and some are genuine mails and is tested if the spam mails are identified or not. Also, it is made sure that the filter learns from the new incoming spam mails. The filter effectively ensures that user’s inbox is not flocked with spam mails.

2.3 Functionalities

Spam filter developed could be of great use to any users using email services. Personalization in filtering out spam by the user can be done by the prototype developed. During the development of the prototype, an interface displaying mails from the user email account is developed. Interface is user friendly with different modules implemented for identifying spam. Support using database and server is observed during the functioning of the filter. The filter is
implemented followed by testing and results are evaluated. Outcome of the project are analyzed based on the effectiveness in filtering out spam.

3. PROPOSED SYSTEM

3.1 Open source email spam filter

SpamKiller is implemented to work effectively on users system to remove spam. It is developed to work with:

Reliability: No matter how many spam mails flock the mailbox, the spam filter would effectively organize the incoming mails by identifying whether a mail is spam or not. At any point, the working of the spam filter would be reliable.

False positives: Avoidance of false positives is of high priority. False positive is a situation where a genuine mail is identified as spam and is directed to the spam folder. Spam filter should be effective enough in differentiating between genuine mails and spam mails. False positives avoidance would increase the accuracy of the spam filter.

3.2 Framework

Spamkiller is developed taking into consideration some of the unique features brought in from several famous spam filters that are already in use. Features that could be effective in identifying spam are implemented in a single spam filter. Spamkiller is developed showing the implementation. An email interface is developed to make it more user-friendly. Check for spam is done in background while importing the mails from mail account to the spam checker. A web based tool is developed where in an email interface displays the options regarding spam check.
For incoming mails, the subject, body etc., of the message is parsed and is checked against a maintained database for spam recognition. The result of the analysis of the mail is also displayed for user convenience. Also, many layers of check is performed by the features implemented thus increasing the efficiency of working of SpamKiller. Options to check the attachments as well the images make it more accurate in identifying spam mails. Due to layers of checking, genuine mails are not identified as spam and spam mails may not remain unrecognized through SpamKiller. Finally, the email interface and spam identification are implemented in the web application and the results are presented in the paper.

3.3 Step by step process of project development:

A brief step by step process of spam filter development is as follows:

1. Acquiring information about unique features from different spam filters available online.
2. Developing designs of a spam filter with all the acquired features.
3. Connection to server established using IMAP and mails brought in using POP3 mechanisms.
4. MySql database created for incoming mails and spam check procedures.
5. PHP based code for spam check implementation.
6. Layer wise sequential check over all the incoming mails.
7. PHP and HTML based email user interface developed.
3.4 System Requirements

The basic resources required in for developing and executing the project are as follows:

**Hardware Requirements**

Processor      : Pentium IV

Hard disk      : 40 GB

Monitor        : 15 VGA color

RAM            : 1GB

**Software Requirements**

Operation system    : Windows 7 Professional

Frontend          : PHP

Backend           : MySQL
3.5. System Architecture:

Figure 2 represents the architecture of SpamKiller that is implemented. The modules of the spam filter are also represented in the proposed architecture:

Figure 2: Proposed architecture of the SpamKiller representing its modules
3.6 Flow of execution:

Figure 3 represents the flow of execution of SpamKiller. Decision making involving in the project is also clearly depicted.

![Flowchart indicating flow of execution of SpamKiller](image)

Figure 3: Flowchart indicating flow of execution of SpamKiller
3.7 Proposed mechanism

Implementation of SpamKiller involves several phases of development. First phase of the development involved acquiring of unique features from several spam filters in use these days. SpamKiller with several features is implemented. Module based development is used to add the features. Each module is individually tested to check whether or not spam is identified. A user friendly email interface is developed with all the options to create email, send email etc.

4. FUNCTIONALITY OF APPLICATION

The overall functionality of SpamKiller is divided into the number of features implemented. A modular approach of development is observed. Implementation of each feature is considered as a module. Hence, there are several modules each for white list, blacklist, grey listing, detection of malware URL’s, implementing Bayesian method in filtering and ability to parse image spam.

4.1 Modular description:

There are several modules implemented in the development of SpamKiller.

1. Maintaining White lists
2. Maintaining Blacklists
3. Grey listing
4. Bayesian filter
5. Detect phishing/ malware URL’s
6. User friendly email interface
• *Maintaining White lists:* MySQL database is created to support the maintenance of white lists. White lists contain list of email addresses that are considered genuine by the user. These could be customized by the user as per preferences. Whenever a new name is added to the white list, an entry is created in the database and for every incoming mail; the sender information is checked against white list. If the mail is from white list entry it is directed to inbox without further checking.

Figure 4 displays options like White list, Black List Email, Black List Word, Black List Phrase, Blacklist URL is shown. Clicking on the option would display the facility to add and delete entries.

![SpamKiller user interface ‘Settings’ option screen.](image)
Figure 5 and Figure 6 show that email ids added to this list are considered genuine and any mails from these users are directed to inbox without further checking.

Figure 5: Screen displaying the entries added into the white list.
• **Maintaining Blacklists:** Blacklists is another feature also implemented in the filter. Several of the users are known to spread spam mails and hence are blacklisted. All the user id’s added to the blacklist are maintained in the database. For every incoming mail, the sender’s information is matched against the list of ids in the database. If found in blacklist without any further checking the mail is directed to spam folder directly. Any number of mail ids could be added to the blacklist. Ids in blacklist can be customized by the user. Choice of ids in the blacklist can be as per user preferences.
In Figure 7 and Figure 8, any mail from these ids is directed to the spam folder without further checking.

**Figure 7:** List of email ids that are added into the blacklist.
Figure 8: Operation of adding a new entry to the blacklist is shown.

- **Maintaining blacklist words /phrases:** Content based parsing is implemented in the project. To support content based parsing, spam words that have high probability of occurrence are stored in the database. Many words can be inserted into the database. Mails with certain words may be sent repetitively and hence users can add these words into the database. Any time later if a mail with these words comes in, the mail is identified as spam and is directed to spam folder. Words can be added and deleted from the database.
Statements in subject or body can also be used to identify the spam mail. Hence, these phrases as a whole can be inserted into the database. Blacklisted phrases are checked for in the body of the incoming message and if identified the mail is directed to spam folder. Figure 9 represents blacklist words added to database and Figure 10 represents mails with blacklisted words.

Figure 9: Words added to the blacklist are displayed.
Figure 10: Adding new entry to the blacklist words operation is displayed.
In Figure 11, phrases with probability of occurrence in spam mails can be added to blacklisted phrases. For mails having subject or body content with these phrases are directed to the spam folder.

![List of phrases added to Black list of phrases is shown.](image)

- **Greylisting**: Greylisting is a nascent technique that has come very recently into implementation. Database support is taken in this methodology. For every incoming mail, if the sender mail id is not present in the white list, and if the mail sent by the sender is identified as spam then the mail id is added to another list called the greylist. If another spam mail is sent by the same sender for the second time, the sender is then added into the blacklist thus blocking any further incoming mails into the inbox. Greylist mails are directed to the spam folder. When added to the blacklist, the mail sender is checked in the blacklist and if available the mails are sent to spam thus blocking majority of spam from
entering the user inbox. Hence, first time spam senders are maintained in the grey list and then are added to the blacklist if a spam mail comes in from the same sender for the second time.

In Figure 12, grey listing is done automatically by the filter and hence only deletion of entries can be done by the user.

![Grey list maintenance in the ‘Spam’ option of the user interface.](image-url)
Figure 13 shows new id that is not present in the white list and when the observed mail is a spam then the id is entered into the grey list by the spam filter.

Figure 13: Email Ids of first time spam mail senders maintained in the grey list.

- **Bayesian filtering:** Bayesian filtering is an adaptive enabling method implemented in the spam filter. Bayesian filtering makes use of Bayes classifier where in tokens in the mail are considered to find probability of spam mails. Bayesian filtering helps in training the filter about new spam mails when content based spam filtering is implemented. Bayesian filter considers all the spam mails and parses the body and subject of the mails. Text in
the spam mails is divided into tokens of some length and the words are inserted into the database. Mails that are spam are used to learn further about the probability of occurrence of certain words. The count is incremented if the same word appears in other spam mails. The more the match with the words, the more probability the mail is spam.

Figure 14 shows that by clicking on the ‘Bayesian filter’ option, spam filter parses all the spam mails present, divides the text in mails in to tokens and stores them in the database. Thus, adaption is possible to the filter.

**Figure 14:** Bayesian filter option present in the ‘Spam’ folder.
Figure 15 shows that tokens from the spam mails are obtained and maintained. The number of occurrences of the token words are counted and maintained. The more the count as well number of matches with the filtered words, the more the incoming mail probability to be spam.

Figure 15: Bayesian filtering showing tokens and respective count.

- **Identify embedded URL**: URL’s from spam mails can direct user control to phishing websites or download harmful files into the system that could spread Malware. Hence, such URL’s can be blacklisted too. URL’s that are obtained from spam mails can be put under blacklisted URL’s and hence any mail with a match with the URL is sent to the spam folder. Such URL’s are all maintained in the database and the database can be extended as the spam mails grow. Usually when Bayesian filtering is done, the words in
the URL are also stored as tokens and hence a match in the mail directs the mail to spam folder.

Figure 16 shows URL’s that have high probability of appearing in spam mails are added to the black list of URL’s.

![Figure 16: URL’s added to black list of URL’s are displayed.](image-url)
Any number of new URL’s can be added to the black list of URL’s. In Figure 17, currently available entries of URL’s can be modified or deleted from the list. Part of URL’s can also be maintained and even a partial match in the incoming mails is taken into consideration.

Figure 17: Options to edit delete or add new URL’s is shown.

- **Web email interface:** An interface is developed to make the spam filter more users friendly. Interface has a plain and simple layout of the inbox mails showing the subject for brief introduction about the message. If implemented on a server, mails can be sent and received out from this interface. Spam mails or mails in the inbox can be deleted and sent to trash. A separate option for spam is present. Figure 18, 19, 20, 21, 22, 23, 24, 25 displays different features implemented is displayed in spam folder.
Figure 18: The inbox of the interface developed is shown.

Figure 19: The ‘Spam’ folder of the interface developed is shown.
Figure 20: Spam statistics option showing mail statistics.
Figure 21: Blacklisted data along with evaluated spam percentage.

Figure 22: Display of Spam mails with brief subject and sender information.
In Figure 23, to the right corner is the display of percentage by which the mail is identified as spam.

Figure 23: Display of spam mails when opened.

Figure 24: The display screen for ‘Settings’ option of SpamKiller.
Figure 25: ‘Trash’ screen that contains any deleted mails is shown.
5. TESTING AND EVALUATION

Testing is the process of running the application to identify any presence of bugs. Testing is done to check whether or not the program executes as required. Testing is also done to validate the running of application. Different types of testing are performed to know any occurrences of bugs and repair them to avoid further problems.

This project is tested in different ways to make sure that all the modules of the project work as expected. Test cases are designed to check every individual functionality and overall functionality of the project. Several types of testing are applied on this project. They are:

1. **Unit testing**: All the small individual units of the project are tested to verify if the functionality of each unit is as expected or not. This is also called module testing as each module of the application is separately tested for expected output.

2. **Integration testing**: Individually testing the units would never be enough as functionality of one unit may cause a hindrance for the functionality of other. Therefore, testing is also done after all the working modules of the application are integrated. Testing is done to know if all the units as a whole give out the required output or not.

3. **Validation testing**: Integration testing is generally followed by validation testing. Validation testing is done to check if the project executes in the least expected way required. Validating of the working of the project is observed.
4. **Output testing**: No project is successfully completed if the output of the project is not as expected. Hence, testing is done to check if the project succeeds to meet the required criteria and gives out a satisfactory expected output.

Following test cases are used to test the working of the project:

**Test case 1**: Verifying if the user can successfully login with Gmail or islander email credentials.

- Any user with valid Gmail email account credentials can login to the system.
- Any user with islander.tamucc.edu email credential can login to the system.
- Incorrect username or password given would prompt to enter the correct credentials.
- Correct credentials would direct to the inbox page with all mails but no spam mails.

Gmail or Islander email credentials can be used. In figure 26, display screen where valid credentials would direct to inbox, if invalid, prompt to enter valid credentials is displayed.

![Figure 26: Login screen where the user’s credentials need to be entered.](image.png)
**Test case 2:** Testing the white list feature functionality:

- Any mail id can be added to the white list.
- Any number of email ids can be added, edited or deleted.
- Mails from ids in white list are directly sent to inbox.

Figure 27 shows the list of ids added to the whitelist.

![Figure 27: Adding email addresses to the white list is shown.](image-url)
Figure 28 shows the mails from white listed ids. Any mails from ids in white list are shown directly in the inbox without any checking.

Test case 3: Testing the blacklist functionality:

- Figure 29 shows mail id added to the blacklist.
- Any number of email ids can be added, edited or deleted.
- Figure 30 shows mails from ids in blacklist that are directly sent to spam folder.
- Users assumed to spread spam are added to blacklist.
Figure 29: Email address added to the black list is shown.

Figure 30: Mail received from ids listed in black list shown in spam emails.
**Test case 4:** Testing blacklisted word:

- Words can be blacklisted. Words that appear in spam can be added here.
- Any number of words can be added, deleted and edited.
- Mails with blacklisted words are identified as spam and directed to spam folder.
- Common words in repetitive unwanted mails can also be added to blacklisted words.

In Figure 31, ‘spamfiltertest2@gmail.com’ id is removed from the white list. Any emails with presence of blacklisted words are directed to spam folder.

![SpamFilter](image)

*Figure 31: Words added to Black list word is shown.*
In Figure 32, as ‘spamfiltertest2@gmail.com’ is removed from white list it is considered as an unknown mail id.

**Figure 32:** Email received with the word listed in Black list words shown in spam emails.
Test case 5: Testing blacklist phrase:

- Not only words, phrases consisting group of words can be blacklisted.
- Any number of phrases can be blacklisted by the user.
- Phrases common in spam mails can be added in here.
- Mails with blacklisted phrase text are redirected to spam folder.

In Figure 33, phrases can be edited, deleted or added in here. In Figure 34, email containing these phrases will be directed to spam emails.
Figure 34: Mails with the phrases listed in black list phrase are shown in spam emails.

Test case 6: Testing blacklisted URL:

- URL’s of phishing sites and malware spreading sites can be added.
- Spam mails have common URL’s to unwanted sites.
- URL’s can be added, edited and deleted in the blacklist.
- Figure 36 shows mails with such URL’s in the spam folder.
In Figure 35, ‘spamfiltertest3@gmail.com’ is removed from blacklist.

Figure 35: New URL added to the blacklist of URL.

Figure 36: No mail with blacklisted URL is observed in the inbox.
**Test case 7:** Testing Grey listing:

- Spam mail from a new unknown sender is added automatically to grey list.
- Entries in grey list can be edited and deleted only.
- First spam mail from new sender is not even added to spam mails.
- Spam filter checks grey list for all the incoming mails.

In Figure 37, while testing for blacklist phrase, a mail is received from id ‘spamfiltertest3@gmail.com’ that is identified as spam because of presence of blacklisted phrase.

![SpamFilter interface](image)

**Figure 37:** First spam mail from new id is added automatically in grey list.
In Figure 38, as first spam mail from new id is identified, id added to grey list.

Figure 38: As first spam mail from new id, mail not added to ‘spam mails’.

- A second spam is received from the same mail id.

In Figure 39, a second spam mail from id ‘spamfiltertest3@gmail.com’ has come in. As a result of which previously available genuine mail in inbox from that id is also redirected to spam folder.
Figure 39: Mails from grey listed sender moved out of inbox after second spam mail.

In Figure 40, on receiving second spam mail from same mail id, both the mails are directed to spam folder.

Figure 40: First second mail from new spam sender is shown in ‘spam mails’.
In Figure 41, on receiving second spam mail from same email id, the mail id is automatically added to blacklist and removed from grey list. Any further mails from the id are directly sent to spam mails.

![SpamFilter interface](image)

**Figure 41:** Grey listed mail id added to blacklist on receiving second spam mail.

**Test case 8:** Testing spam statistic option

- Based on total number of mails percentage displayed.
- Figure 42 shows Percentage of spam mails obtained and also,
- Percentage of mails obtained in inbox is displayed
Figure 42: Percentage of spam mails and secured mails shown.

In Figure 43, only one mail is in inbox, all other mails are in spam.

Figure 43: Few mails are available in inbox as per statistics.
Test case 9: Testing blacklist in spam options

- History of previously stored blacklist words, phrases, URL’s are maintained.
- Figure 44 shows currently added blacklist phrases, words. Figure 45 shows URL’s that are maintained.
- Mails with presence of any of the blacklisted content are considered as spam.

Figure 44: Currently added blacklist phrases along with previously stored blacklist data.
Test case 10: Testing the mails identified as spam:

- Mails from mail ids present in blacklist are present in spam mails.
- Mails with blacklisted words, blacklisted phrases, blacklisted URL’s are in spam mails.
- Mails identified as spam as a result of grey listing are added to spam mails.

Figure 46 displays the spam mails that are identified. Figure 47 shows the id that is blacklisted on receiving second spam mail from same user id.
Figure 46: Spam mails identified are shown in ‘Spam’ folder.

Figure 47: Here, all the ids present in blacklist, as well ids added because of grey listing are displayed.
Test case 11: Testing Bayesian filter option:

- Figure 48 shows all spam mails present that are parsed and words are identified in mails.
- Fixed length tokens of all the text in spam mails are maintained.
- All incoming mails checked against words present and count increases if match found.

Figure 48: Spam mails that are parsed during Bayesian filtering are displayed.
Figure 49 shows text in spam mails that is parsed to maintain fixed length tokens. Count indicated the number of occurrences in spam mails. And percentage is displayed indicating the growth as count increases of an individual token.

![Image of a table showing Bayesian filter words maintenance with respective count and percentage.]

Figure 49: Bayesian filter words maintenance with respective count and percentage.
6. CONCLUSION AND FUTURE WORK

Spam mails have been a never ending problem for users using email services on internet. Flood of mails with unwanted information are forced into users inbox which need users manual effort to delete. Observed wastage of bandwidth is high in circulating the spam mails. Also, organizations spend huge amount in stopping spam from interfering with organization email services. Spam filters can be of great use in filtering out spam. Several spam filters with powerful features are usually available only at a cost. Hence, in this project an open source email spam filter called SpamKiller is implemented that could be made available online with no cost to use the features provided. The features implemented are grabbed from several other spam filters that are available at a cost. For the spam filter to execute; the server connection to the email provider is established using IMAP protocol and mails are fetched using POP3 protocol. Several layers of scrutiny are followed to stop the spam mails from reaching inbox of the user. Grey listing and Bayesian filtering help in stopping spam and also training the spam filter in accordance with the new spam mails trend. All the incoming messages are parsed for presence of any spam content and based on the matched content; decision about spam mails is taken.

SpamKiller is tested to have a satisfied working of all the features. Features are implemented as modules and hence every module is tested for expected working. Results of testing show that SpamKiller works successfully in filtering out most of the spam mails.

A huge scope for enhancement is possible in this project. Many new features can be added to the current version of the spam filter. The filter can be implemented on a server to enable the email functionality of the project. Efforts can be put in the direction to decrement the time taken to train the spam filter. Filter can be enhanced to support different other types of spam besides text spam only.
BIBLIOGRAPHY


