Mitigating Spam Emails Menace Using Hybrid Spam Filtering Approach

Stanlee Nagaroor and G.A. Patil

Abstract Spam is a known problem to email users. Sending spam emails is one of the easiest form of advertising and hence a useful medium of communication like email is abused by spammer to send junk email. Most antispam solution focus only on analysing text of the body of email messages for detecting spam emails. We have developed a spam filter that separates spam from non-spam (ham) emails by analysing header, URL, body, and attachments. Header and URL are checked against rules and text of body and attachments are checked by Bayesian classifier and Apriori algorithm. Only (*.rtf, *.txt, *.docx, *.doc,*.pdf) attachment files are examined. The experimental results reveal that checking attachments of emails played significant role in spam detection and hence attachment checks should be extended to more file types for better spam detection.

Keywords Email · Spam · Bayesian · Apriori

1 Introduction

Spam email can be defined as unsolicited or bulk mail sent for the purpose of advertising product or for other malicious intent. According to commtouch [1] Internet threats trend report, 97.4 billion spam emails were sent each day during the first quarter of 2013. Most of the spam emails are designed to solicit money from the recipients and to achieve this they offer products that claims to miraculously cure health problems like diabetes, obesity, hair fall, etc. Spammers also know that humans are greedy by nature and hence they come up with get-rich schemes which claims to make you rich in no time. Spam emails also contain sexual content that

S. Nagaroor (✉) · G.A. Patil
Department of Computer Science & Engineering, D. Y. Patil College of Engineering & Technology, Kolhapur, India
e-mail: stanlee.n@gmail.com

G.A. Patil
e-mail: gasunikita@yahoo.com

N.R. Shetty et al. (eds.), *Emerging Research in Computing, Information, Communication and Applications*, https://doi.org/10.1007/978-981-10-4741-1_20

involve pornography or promote some drugs or products that could enhance sexual performance. Some of the countries have enacted laws like the CAN-SPAM act, The CAN-SPAM Act is a law that sets the rules for sending of commercial email and has provision of penalties for violations. However all countries do not have an antispam law. Even if such law exists it is not implemented effectively. According to spamhaus [2], United States, China, and Russia are top three generator of spam emails.

2 Related Work

There are many approaches to deal with the problem of spam and different researchers have come up different solutions and different algorithms to detect spam. Most of the work focuses on analysing the content (body of email) message. These include using algorithm like naïve Bayes [3, 4] for classifying email messages into spam and non-spam emails. Some of the algorithms that are commonly used for classification of emails are SVM [5], RF [6], K means cluster [7], K nearest neighbour [8].

Some researchers focused on extraction of features from header [9, 10, 11] to find if it can be used for spam detection. Researchers have also focused on using digest [12, 13] of emails to detect spam emails. Structural features [14] of emails are also used for detection of spam emails. A system was also developed for detecting phishing URL [15] based on Lexical-based features, Keyword-based features, Reputation-based features, and Search Engine based features. Other techniques use block list like Domain Block List (DBL) the block list of the spamhaus [16] project is used by several organisations and email service providers. To best of my knowledge, researchers have written too little on attachments spam and techniques for detection of attachment spam is not made public by those software companies that claim to detect attachment spam.

3 Methodology

Spammers want readers to respond to their mails, hence most of the spammers provide URL in the message body that direct to spam website. Hence examining URL is important in spam detection. Also spammers use fake headers to hide their identity or to pretend to be someone else. Using black list or white list against sender email address will not give correct results, hence other header fields should also be examined. The proposed spam filter will examine other fields in headers too.

Emails from mail server are read using IMAP protocol. The spam filter examines headers and URLs of emails for classification. Only some emails are classified by header and URL checks alone. Emails which cannot be classified by header or URL checks undergo body and attachment analysis where Bayesian classifier is used for classification. The emails which are classified as non-spam (ham) by Bayesian classifier is further examined for the presence of spam-associated words (generated by Apriori algorithm) which will help to detect spam emails not detected by

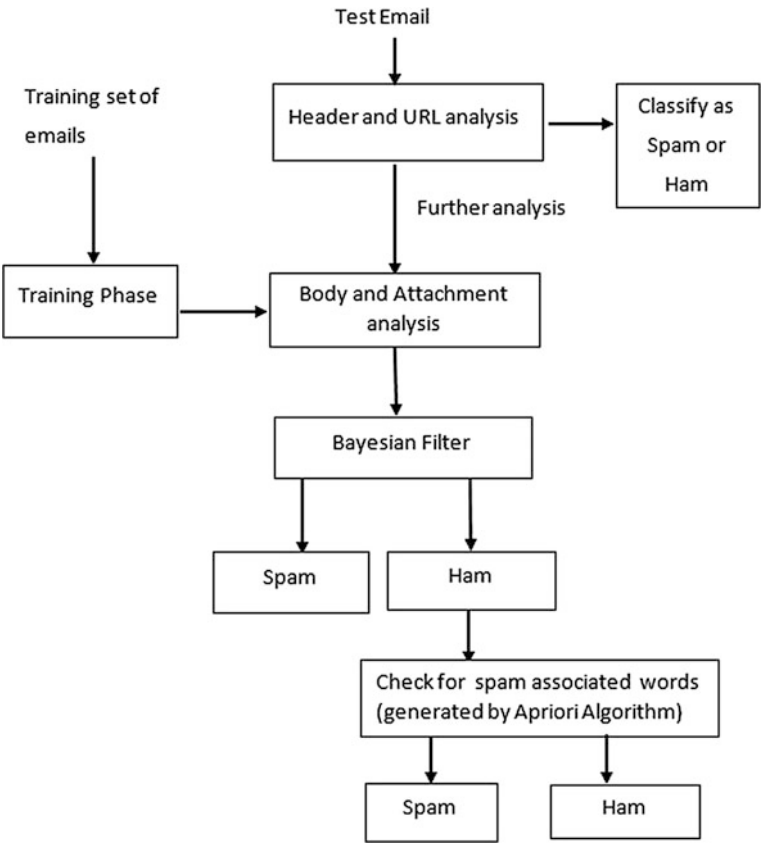

Fig. 1 Architecture of spam filter

Bayesian classifier. Emails that are classified as ham by Bayesian classifier and if it does not have spam associated words it will be considered as ham (Fig. 1).

We can examine body and attachments only after training the spam filter using training dataset of spam and ham emails. The knowledge gained by the filter during the training phase will be used for analysis of body and attachment in the testing phase.

Section 3.1 gives description of training of our spam filter while Sect. 3.2 describes the process of detection of spam emails by analysing different parts of email messages.

3.1 Training of Spam Filter

Bayesian classifier must be trained to distinguish spam emails from non-spam (ham) emails. To train the classifier we must provide two datasets (spam dataset and ham dataset). Bayesian classifier scans through the training sets of ham and spam

emails and takes note of the words occurring in the datasets and also the proportion in which these words appear in dataset. This knowledge helps to assign proper probabilities to words.

Each email in spam or ham dataset goes through following steps:

I. Removal of html tags.
II. Removal of stop words.
III. Stemming of words.
IV. Storing words in database and calculating probability.

To reduce computation cost and unnecessary space overhead, stop words and html tags are removed from the emails. To correctly take in account the occurrence of a particular word in spam and ham message the words are stemmed, i.e., words like waiting, waited are reduced to their root form wait.

The details of each steps in training are given below

I. Removal of HTML tags and entities

Regular expression are used to replace or remove HTML tags. HTML tags like <script></script>, <style></style> are completely removed along with the text that falls between the opening and closing tags. Some tags are replaced by appropriate text if a particular pattern matches, for example,
 tag is replaced with "\n". Regular expressions are also written to handle entities, special symbols, for example, < is replaced with '<'.

II. Removal of stop words

Stop words are those words that do not help in classification and hence are removed. Examples of stop words include articles, conjunction, pronouns, and preposition. If the email contains stop word it is removed from the email as a part of pre-processing step. Removal of these words reduce computational and space overhead.

III. Stemming of words

After removal of HTML tags and stop word from the text, the words that contains only numbers are removed at this stage. Also words are reduced to their root form using Porter stemming algorithm.

IV. Storing words in database

After the messages in spam dataset passes through above three steps. The remaining words (tokens) are stored in spam table of database. If a single email contains duplicate words, it is removed before storing it in spam table of database. By removing duplicates the frequency corresponding to a word will indicate how many spam emails contains that word for example, if word 'free' occurs four times in spam email one and three times in spam email two. The spam table will contain entry <free,2> instead of <free,7> indicating 'free' word is seen in two spam mails.

Similarly, emails from ham dataset is processed by removing HTML tags, removing stop words and applying stemming to the word. The remaining words are

stored in ham table of database. If a single email contains duplicate words it is removed before storing it in ham table of database.

Probability that a word is spam is calculated using formula,

$$P\left(\frac{S}{W}\right) = \frac{P\left(\frac{W}{S}\right) . p(S)}{P\left(\frac{W}{S}\right)p(S) + P\left(\frac{W}{H}\right)p(H)} \quad (1)$$

where,

1) *P(S/W)* is the probability that a message is a spam, knowing that the word W is in it.
2) *P(S)* is the overall probability that any given message is spam.
3) *P(W/S)* is the probability that the word W appears in spam messages.
4) *P(H)* is the overall probability that any given message is not spam (ham).
5) *P(W/H)* is the probability that the word W appears in ham messages.

There is no prior reason for an incoming mail to be considered as spam or ham so we assign equal probabilities for *P(S)* and *P(H)*, hence *P(S)* = *P(H)* = 0.5. The above Eq. (1) becomes

$$P\left(\frac{S}{W}\right) = \frac{p\left(\frac{W}{S}\right)}{P\left(\frac{W}{S}\right) + P\left(\frac{W}{H}\right)} \quad (2)$$

Apriori Algorithm is used to generate sets of spam associated words from training data set of spam emails. To apply this algorithm to find association between spam words, we have to create our transaction database from training dataset of spam emails by selecting spam words from each training email (all spam words in single mail will be considered as a single transaction). Transaction database (T) will contain several such transaction. Then we will use Apriori algorithm to generate sets of spam associated words like {win, lottery}. This indicates that spam words won and lottery are closely associated (or related).

Attachments will be read and text from the attachments will be analysed using Bayesian and Apriori algorithm for classifying emails.

3.2 Detection of Spam Emails

Following checks are performed on headers of emails to detect spam emails:

1. SPF checks: Sender Policy Framework (SPF) is an email validation technique used to detect email spoofing by providing a mechanism to check that incoming email from a domain comes from a host authorised by that domain's administrators. If an email claims to come from a certain domain but the IP address of sender is not in SPF record published by domain's administrator then mark that email as spam.

2. MX record checks on the 'from' field of the message: The reason for performing MX record check is that some spammers use nonexistent domain names in 'from' field of header. To check whether the domain in the 'from' field has a mail exchange record we send a query to DNS server asking for MX records for the domain in the 'from' field. If we get MX record details from DNS server then domain name in 'from' field exists otherwise it means that nonexistent domain is used in 'from' field and hence the message is marked as spam.
3. Check if an email is a real reply email: If the subject contains 'Re:' check if the 'In-Reply-To' header is present. In a genuine reply message 'In-Reply-To' header contains message id of email for which this message is a reply. Spammers sometimes use 'Re:' in their subject but it is a fake reply message. Hence the 'In-reply-to' field is not present in the header of spam email or it does not contain value in it.

   In such cases mark that email as spam.

4. Detect spam pattern in subject. Regular expressions are used to detect presence of spam like pattern in subject.
5. If an email passes SPF checks and domain in 'from' field and 'return-path' matches, then the domain is compared against those in white list. If the domain is present in white list, mark email as ham.
6. Check headers against black list: check address in 'received' fields and 'from' field against black list. If the domain is present in black list mark that email as spam.

If content type of email body is html then first an Html Document object is created and then descendants 'form' nodes, 'anchor' nodes and 'area' nodes are extracted. If URL is present in form's action attribute or anchor's href or area's href attribute, then these URL's are extracted and following checks are done on URL.

I. Domain name from the URL is compared against black list. If the domain name is present in the black list mark email as spam.
II. Check if text and href of anchor contains URL and whether these URL's point to different domains.

For example, <ahref=http://www.scamsite.com>www.paypal.com. Spammers use such technique to direct user's to phishing site. If the URL in href and text point to different domain's mark that email as spam.

The text from body of email will be extracted. Attachments of email are read as sequence of bytes, then the bytes are decoded and formatting information is removed to get data in plain text. Different file types have different formatting info. Only (*.rtf, *.txt, *.docx, *.pdf) attachment files are examined. Attachments are

Fig. 2 Conversion of bytes in attachment to plain text

read as sequence of bytes and it is decoded to get text with formatting information and latter the formatting information is removed to get plain text as shown in Fig. 2.

The text from body or attachment will be preprocessed (tokenization, stop word removal, html tags removal, and stemming). The description of preprocessing steps are given in Sect. 3.1.

Spamicity of words in email will be calculated by formula (2).

Probability (p) that an Email (E) containing words $\{W_1,W_2,W_3,\ldots W_n\}$ (in body or attachment) is spam is given by formula (2).

$$P = \frac{P\left(\frac{S}{W1}\right)P\left(\frac{S}{W2}\right)P\left(\frac{S}{W3}\right)\ldots P\left(\frac{S}{Wn}\right)}{P\left(\frac{S}{W1}\right)P\left(\frac{S}{W2}\right)\ldots P\left(\frac{S}{Wn}\right)\ldots + \left(1 - P\left(\frac{S}{W1}\right)\right)\left(1 - P\left(\frac{S}{W2}\right)\right)\ldots\left(1 - P\left(\frac{S}{Wn}\right)\right)} \quad (3)$$

where,

$P\left(\frac{S}{W1}\right), P\left(\frac{S}{W2}\right), P\left(\frac{S}{W3}\right), \ldots, P\left(\frac{S}{Wn}\right)$, are spamicity of words $W_1,W_2,W_3,\ldots W_n$.

Emails having value of p above a threshold are considered as spam. Rest of the emails are considered ham by Bayesian classifier.

The emails that are considered as ham by Bayesian classifier are further examined for the presence of spam words association. The spam words from emails are retrieved and checked against spam association rules (or sets of spam associated words, generated by Apriori algorithm). If association is present then email is marked as spam else it is marked as ham.

4 Experimental Results

We use the following parameters to measure performance of spam filter:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where,

TP = True positive, i.e., spam emails correctly identified as spam
TN = True Negative, i.e., ham emails correctly identified as ham
FN = False Negative, i.e., spam emails that are not correctly identified
FP = False Positive, i.e., ham emails that are not correctly identified

Total Emails in dataset used for Testing: 250, Ham emails in dataset: 200, Spam emails in dataset: 50. The results given by our system is shown in Table 1.

Table 1 Correctly classified and incorrectly classified emails

	Correctly classified	Incorrectly classified
Ham	193 (TN)	7 (FP)
Spam	45 (TP)	5 (FN)

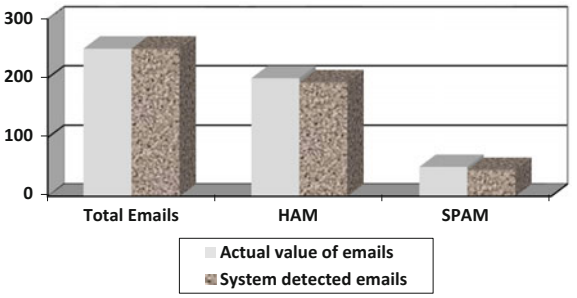

Fig. 3 Comparison of actual value and system generated result

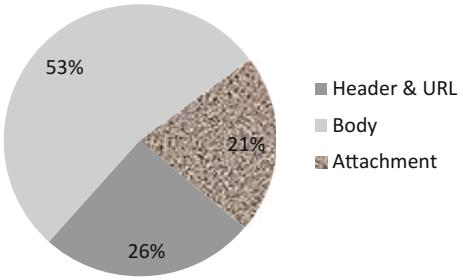

Fig. 4 Distribution of spam emails detected by analysing different parts of email messages

$$\text{Accuracy} = (193 + 45)/250 = 0.952 = 95.2\%$$
$$\text{Precision} = (45)/(45 + 7) = 0.865 = 86.5\%$$
$$\text{Recall} = 45/(45 + 5) = 0.9 = 90\%$$

The column chart (Fig. 3) shows the number of actual ham and spam emails in dataset and ham and spam emails correctly detected by the system.

The pie chart shown in Fig. 4 shows how each part of email messages helped in detection of spam emails. This diagram shows that 53% of spam emails were detected by analysing the body, 21% of the spam emails were detected by analysing attachments and remaining percent of spam emails were detected by header and URL checks.

5 Conclusion and Future Work

Spam emails are unwanted emails and needs to be weeded out if we have to reinstate confidence of the people on this useful medium of communication. This paper discussed the different approaches used by researchers to deal with the

problem of spam. This paper gives details on our approach to solve the problem by performing checks on header and URL using rules and analysing the text of body and attachments using Bayesian classifier and Apriori algorithm. Our future work will be on extending our work of analysing attachments on other file types and checking their impact on overall spam detection of spam filter.

References

1. "Internet threats trend report", Commtouch, April 2013.
2. https://www.spamhaus.org/statistics/countries, March 2016.
3. G. Bhagyashri and H. Pratap "Auto emails classification using Bayesian filter", International Journal of Advanced technology & Engineering Research, vol 3, Issue 4, July 2013.
4. Smera Rockey and Rekha Sunny "A hybrid spam filtering technique using Bayesian spam filters and aritificial Immunity Spam Filters" International Journal of engineering research and technology", vol 3, Issue 5, May 2014.
5. Sushama Chouhan, "Behavior Analysis of SVM Based Spam Filtering Using Various Kernel Functions and Data Representations", IJERT, vol. 2, Issue 9, September – 2013.
6. B.Gaikwad and P. Halkarnikar, "Random Forest Technique for E-mail Classification", International Journal of Scientific & Engineering Research, vol 5, Issue 3, March-2014
7. Nadir Omer, Othman Ibrahim and Waheeb "An improved spam email classification mechanism using K means clustering", vol 60, No 3, February 2014.
8. Ali Aski "A Proposed Algorithm for Spam Filtering Emails by hash table approach", International Research Journal of Applied and Basic Sciences vol.4 (9), 2436–2441, 2013.
9. Omar Al-Jarrah, Ismail Khater and Basheer Al-Duwairi "Identifying Potentially Useful Email Header Features for Email Spam Filtering", ICDS 2012: The Sixth International Conference on Digital Society.
10. Alberto Treviño, J. J. Ekstrom "Spam Filtering Through Header Relay Detection", Brigham Young University.
11. Fernando Sanchez,Zhenhai Duan,Yingfei Dong "Understanding Forgery Properties of Spam Delivery Paths".
12. E. Damini,S. De Capitani di Vimercati and P.Samarati "An Open Digest-based Technique for Spam Detection", DTI - University di Milano - 26013 Crema, Italy.
13. E. Damiani, S. De Capitani di Vimercati, S. Paraboschi, and P. Samarati. "An open digest-based technique for spam detection". In Proceedings of The 2004 International Workshop on Security in Parallel and Distributed Systems, San Francisco, CA, USA, September 2004.
14. Sarju,Riju Thomas and Emilin shyni "Spam email detection using structural featues", International Journal of computer Applications,vol 89, issue 3, March 2014.
15. Ram B. Basnet, Andrew H. Sung, Quingzhong Liu "Learning to detect phishing URL" IJRET: International Journal of Research in Engineering and Technology.
16. https://www.spamhaus.org/
17. G.SenthilKumar, S.Bhaskar and M.Rajendran "Online Message categorization using Apriori algorithm", International journal of computer trends and technology, June 2011.