A Case Study of User-Level Spam Filtering

Kamini (Simi) Bajaj and Josef Pieprzyk
Department of Computing, Macquarie University
North Ryde, NSW, Australia
k.bajaj@uws.edu.au and josef.pieprzyk@mq.edu.au

Abstract

There are number of Anti-Spam filters that have reduced the amount of email spam in the inbox but the problem still continues as the spammers circumvent these techniques. The problems need to be addressed from different aspects. Major problem for instance arises when these anti-spam techniques misjudge or misclassify legitimate emails as spam (false positive); or fail to deliver or block spam on the SMTP server (false negative); thus causing a staggering cost in loss of time, effort and finance. Though false positive are very harmful loss of important information for the user, false negatives defeat the purpose of the spam filtering. This paper makes an effort in proposing another aspect to address this problem. It discusses some of these anti-spam techniques, especially the filtering technological endorsements designed to prevent spam to entrench their capability enhancements, as well as analytical recommendations that will be subject to further research. Apart from applying anti-spam techniques, training of Spam control tool with relevant user preferences can reduce the chances of false positives, false negatives and spam email that land in the inbox. We identify the need for training the filter with domain specific data. This paper shows the decline in false negatives via results of a case study on training the Spam Bayes tool with carefully collected domain specific user preferred dataset for over a period of 12 months.[1]

Keywords: Spam Bayes, training email filters, content filtering, false negatives, user level spam filtering, email spam control

1 Introduction

The most common form of communication these days, in organisations especially and for consumers is email. In 2013, there are 929 million mailboxes for business email accounts and it is expected to grow and reach over 1.1 billion in 2017(Levenstein, 2013).. Also in 2013, the majority of email traffic comes from the business emails, which accounts for 100 billion emails per day. Majority of business communication happens over email (Levenstein, 2013). Since emails are so rigorously used, they come with problems. The problem being focused on in this paper is email spam. Spam has been a nuisance for everyone who sends or receives messages using computer, tablet or smartphones from last decade. The Word which is used for a while to describe unsolicited email was originally the brand name for Hormel Foods, maker of the canned "Shoulder Pork and hAM"/"SPiced hAM" luncheon meat. By tracing the history, SPAM originated in 1970's as a repetitive advertising message that was sent to a large number of recipients with or without subscribing for the advertising message. SPAM in early 1980's was an innovative way of sending information to large cross sections of people, however today it has become a menace. Spam itself is controversial as the same message from originator could be the means & ways of advertising the product/services and for the other it is unwanted message, 'a nuisance'. Overall, spam has become one of the major social issues; it is abuse of the electronic messaging system which promotes information, products or services that are not asked for. This abuse is done in various ways for example, turning a machine into a relay for spam, a staging ground to attack other systems, or a spy capturing your bank account and credit card information--or all three(Ford and Spattord, 2007)

According to spam statistics for recent years, the percentage of email sent over the Internet has increased from 36% in 2002(Clifford et al., 2003) , 45% in 2003 to 64% in 2004(Jung and Emil, 2004) [13] to 80% in 2006(Siponen and Stucke, 2006, Leavitt, 2007, Jaeyeon and Emil, 2004), 92% in 2009(J. and T., 2008) and 95% in 2010 (Gina, 2010) out of which 55% came from 10 countries, such as, India, Brazil, Russia, Ukraine, Romania, South Korea, Vietnam, United States, Kazakhstan, Indonesia, Poland, China, Colombia, Israel and Taiwan. This number went down to 86% in 2011(Namestnikova, 2011) and stayed around 70% in 2012 and 2013. But according to predictions for 2013(Jeff, 2013), phishing to be more prominent and rise in spam related to replica products. This has been confirmed according to the reports for Q1 and Q2 of 2013 by SecureList, which state that there is already a rise in figures of spam and phishing. According to Spam Act compliance and investigations by ACMA, the number of spam complaints received for Jan 2013 was 51525 out of which 50477 were complaints related to emails.

The reasons of existence of Spam are the low cost to send Spam mail and the ease of sending it through various software tools list (Takumi et al., 2007, Guzella and Caminhas, 2009, Deepak and Sandeep, 2005, Dhinakaran et al., 2007). But its existence costs big to the consumers , organisations and the countries as per the figures listed here. According to Ferris Research, who study messaging and content control; the cost of spam mail to organizations in United States was USD 8.9 billion in 2002 with a 12% increase in 2003 to $10 billion and $17 billion in 2005(Ferris_Research, 2007). This cost rose to $100 billion in 2007. In Japan, the amount of

[1] Copyright (c) 2014, Australian Computer Society, Inc. This paper appeared at the 12th Australasian Information Security Conference (AISC-2014), Auckland, New Zealand, January 2014. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 149. Udaya Parampalli and Ian Welch, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

GDP loss was about 500 billion yen in 2004(Takemura and Ebara, 2008). In a press release in 2009, Gartner[2] stated that more than 5 billion US consumers lost money in phishing (type of spam) attacks which was 39.8% more than the year before(Gartner, 2009). The negative impacts of spam are waste of computing resources, loss of productivity of users, denial of service, invasion of privacy , fraud and deception(Ferris_Research, 2007, Nagamalai et al., 2007, du Toit and Kruger, 2012).

Many techniques have been used to control spam (these will be briefly discussed in section 2) and some of them such as content filtering using Bayesian filters have been successful to a large extent but the problem still persists as is evident from the statistics discussed in this section.

It is important to introduce the concepts of false positives and false negatives here before we discuss the problem further. False positive is the legitimate message that is mistakenly identified and marked as spam. False negative is a non-legitimate message not identified as spam; rather spam identification technology marks it as a 'legitimate non-spam message'. Among the various techniques to control spam, content filtering especially using Bayes Theorem is most commonly used and has gained a lot of success. However, we still come across the problem of false positives and false negatives. In this paper, the main focus is to reduce the false negatives in the users' inbox. Though false positives are very harmful but having false positives in the users' inbox defeats the purpose of having spam filtering in place. Hence, we are looking at how the current filtering mechanism can be improved.

Spam email filtering is done at the mail server using various techniques being a mix of listing and content filtering but still the false negatives escape those filters and land into users' inbox. One reason for that could be that the filters used to filter content of the new emails are trained using datasets that contain a collection of generic spam and ham sample messages. An email may be spam for one user and ham for another. We attribute this variation to two parameters. 1. User belonging to a particular industry type; we call it 'Domain' 2. Users' individual preferences. For a user in a particular organisation belonging to domain, a typical collection of email belonging to that domain would be ham in most of the cases. The only variation would be users' individual preferences. In this paper, we have addressed both of the parameters. To do this, we make a hypothesis. Our hypothesis is that false positives in users inbox can be reduced using training dataset that is specific to the domain. To prove that our hypothesis is true, we installed Spam Bayes on the email client of the user and trained it with the said dataset and observed the results for a period of 12 months.

The aim here is to see whether such training can reduce the false-negatives in the inbox on the basis of such training or not. Hence we train the filter with such domain specific user preferred data, then observe the classification done by the filter after the training on incoming new emails and analyse the results to find out if such training reduced the false negatives in the user inbox.

The paper is structured as: Section 2 lists existing solutions to control spam using different techniques. Section 3 involves discussion and explains the need for training filter using Domain specific user preferred data. Since the experiment is done using Spam Bayes, Section 4 gives the background, training model and learning method of Spam Bayes. Section 5 is on Experimenting Spam Bayes which elaborates the experiment conducted for training Spam Bayes tool. This section covers how data was collected, training of Spam Bayes using the data and observation of the results of the training. Finally, section 6 is conclusion and future work.

2 Spam Control Techniques

Anti-spam techniques and methods try to tell apart a spam email from legitimate email. As a typical email consists of few components such as the header, the body and attachments, the algorithms that classify emails may use different features of the mail components to make decision about them. Lot of work has gone into finding solution to spam problem from different dimensions and directions(Islam and Zhou, 2007, Zhang et al., 2012, Xiao-wei and Zhong-feng, 2012, Rajendran and Pandey, 2012, du Toit and Kruger, 2012, Xiao et al., 2010, Wei et al., 2010, So Young and Shin Gak, 2008, Klonowski and Strumiński, 2008, Horie and Neville, 2008, Nhung and Phuong, 2007, McGibney and Botvich, 2007, Liu et al., 2007, Huai-bin et al., 2005, Moon et al., 2004, Wu and Tsai) over the last decade. Various anti-spam solutions are available that have been surveyed by many researchers(Blanzieri and Bryl, 2008, Caruana and Li, 2012, Guzella and Caminhas, 2009, Lai, 2007, Yu and Xu, 2008, Paswan et al., 2012, Nazirova, 2011). Those are blacklists, whitelists, grey lists, content based filtering, feature selection methods, bag-of-words, machine learning techniques such as Naïve Bayes, Support vector machines, artificial neural networks, lazy learning, etc), reputation based techniques, artificial immune systems, protocol based procedures, and so on. (Caruana and Li, 2012) also lists some emerging approaches such as per to peer computing, grid computing, social networks and ontology based semantics along with few other approaches. These solutions can be grouped into various categories such as list based techniques, and filtering techniques; another categorisation can be prevention, detection and reaction techniques(Nakulas et al., 2009). (Paswan et al., 2012) categorises the email spam filtering techniques as origin based spam filtering, content based filtering, feature selection methods, feature extraction methods, and traffic based filtering. The scope of this paper is content based filtering and in specific learning based filters. Hence, we would not go into detail of each of these solutions but limit ourselves to Bayes algorithm.

Spammers are insensitive to the consequences of their activities and need to be dissuaded by being made to pay by the internet service providers for the waste of bandwidth occupied by unwanted spam blocked by the servers. This would be a feasible deterrent to reduce spam. To execute this, all service providers must act in

[2] http://www.gartner.com/it/page.jsp?id=936913

unison and agree to get spammers to pay for spam inconvenience and servers clean up.

3 User Preferred Domain Specific Training of Filter

In this work, we have been able to identity different types of anti–spam techniques exemplified in the use of filters and other characteristic means to deter spammers. Although these anti-spam techniques may be suitable to some users and unsuitable to others, they achieve some level of protection against unwanted email messages. Each of these anti-spam techniques has its unique feature that distinguishes it from the rest of others although none of these are able to perfectly and substantially produce zero false positives and zero false negatives or totally able to stop all real-time and potential spam. The main reason for this is that spammers are always evolving new tricks to deceive the filters. Some well-designed filters (for example, Bayesian Filters) work very well getting success rate as good as 98-99% at certain stages. But this number does not stay the same. Spammers are able to vary these with ease.

In this paper, we are verifying that the filtering mechanisms capability can be enhanced by domain specific training and incorporating user preferences. This enhancement of the filtering capability can increase the performance of the filter. We are trying to build over the fact that, there is no correlation between the receiving user's area of interest and content of spam email. Many researchers have studied the content of Spam messages and various categorisations have been published. One categorisation on the basis of content type is Scams, Adult, Financial, Pharmaceuticals, stock, phishing, diploma, software Malware, gambling, dating and so on(SpiderLabs, 2013). Filters are trained to identify spam on the basis of these categories. But such training by the filter would look for features related to any of those categories generally in any email received. The training dataset would contain features from all of these categories and the features could be confusing for the filter as the spam training could still work but the ham features are anything other than these categories. In fact, the cohort of emails in inboxes belonging users in different domains would be different. For examples, user that belongs to healthcare domain would receive healthcare kind of emails as compared to a user who belongs to real estate.

Same email may be Ham to one user but Spam to another user based upon their preferences. Users in various Domains have difference preference of emails that they would classify as Spam or Ham. For example, an email from a bookseller trying to sell books on Computer Science would be Spam for a Pharmacist. The interesting question that comes out of this is that how do the filters know which emails are Spam and which are ham for the particular user. Of course in some mail clients such as gmail there is an option of user preference setting where user can be given an option to choose the topic area of interest and then the filter can use that information to classify incoming emails accordingly. There is very little work gone into the area of considering specific user preferences while designing anti-spam filters. (Kim et al., 2007) constructs a user preference ontology on the basis of user profile and user actions and trains the filter on the basis of that ontology. (Kim et al., 2006) suggest user action based adaptive learning where they attach weights to Bayesian classification on the basis of user actions. However none of the work address user belonging to a particular domain and their preferences accordingly.

The case of training the filter at the client by the client data is not new; any user who would install SpamBayes would train it with the training data(Meyer and Whateley, 2004). An organisation that uses SpamBayes to filter incoming email would for multiple users would retrain the filter on all received email(Nelson et al., 2009). Training of filter with different feature selection methods is also addressed in (Gomez and Moens, 2010). The novelty in this case is that the data we used for training is the user preferred data carefully collected for a period of 5 years. Second important point that affects the training is that the data collected belongs to a particular user belonging to a specific domain not a general user. This means that we are training the filter that if there is no correlation between the receivers' domain area and the email content, the email is not wanted by the user. User belonging to an educational organisation would have different preferences as compared to a user belonging to a marketing organisation. Different users within an organisation would have different preferences and same message could be classified as spam by different users. The organisational filters cannot take care of such user preferences. Hence, such emails end up in user inboxes as false negatives. The dataset also takes into account such user preferences. The filter is trained on the basis of collection Spam and Ham emails classified by the user belonging to a particular domain.

We made a hypothesis that domain specific user preference training of the filter reduces the false negatives in the user inbox. To justify this hypothesis we chose the spam filtering tool called Spam Bayes, installed it on the outlook mail client and trained it with domain specific user preferred data. The next two sections give details on the background of the tool and the experiments done using the datasets.

4 Spam Bayes-Training Model and Learning Method

This section covers few topics via three major subsections: first is on Bayes Theorem(du Toit and Kruger, 2012, Vu Duc and Truong Nguyen, 2012, Liang and Yu, 2012) which explains how the theorem calculates the probability of occurrence of each word in the document, second is on Spam Bayes which explains what is Spam Bayes, the background, training model and third section is on the learning method in the tool taken from (Nelson et al., 2009)and (Meyer and Whateley, 2004).

4.1 Bayes Theorem

Bayes formula of total probability is

$$P(B \mid A) = \frac{P(B) \, * \, P(A|B)}{P(A)}$$

Applying it here, the probability P(C_i | D) that a document D belongs to a class C_j, can be shown by

$$\frac{P(C_i) * P(D|C_i)}{P(D)} \qquad (1)$$

where: D is an email document to be classified

C_i is one of m classes in class set: $C_1, C_2 \ldots , C_m$

Since there are too many features in the document D, the assumption here is that, probability of each feature in the document D is independent of the context in that it appeared and the location of the features in the document. Probability $P(D|C_i)$ is calculated from the frequency of each feature f_j in the document D:

$P(f_1,f_2,\ldots f_n|C_i) = \prod P(f_j \mid C_i)$

can be re–written as

$$P(C_i|D) = \frac{P(C_i)}{P(D)} \prod P(f_j \mid C_i)$$

From that point, using the principles of probability calculation for single feature or for multiple features by Naive Bayesian algorithm as follows(Vu Duc and Truong Nguyen, 2012):

Let us call the content of each e-mail as: document.

Class Spam email is called 'spam' and

Class Ham email is called 'ham'.

Probability that an email is spam is

$$P(\text{ spam} \mid \text{document}) = \frac{P(\text{ document} \mid \text{spam }) * P(\text{ spam})}{\text{Total}}$$

where total is calculated by

P(document|spam) * P(spam) + P(document | ham) * P(ham)

$P(\text{document} \mid \text{ham}) = \prod P(\text{ feature}_i \mid \text{ham});\ 1<i<n$

$P(\text{document} \mid \text{spam}) = \prod P(\text{ feature}_i \mid \text{spam});\ 1<i<m$

P(spam) = total spam | total messages

P(ham) = total ham | total message

4.2 Spam Bayes Training Model(Nelson et al., 2009))

Spam Bayes born on 19 August 2002 (Meyer and Whateley, 2004) is a freeware tool based on the Bayes theorem which has been successful in controlling spam to a large extent. Since we are using Spam Bayes for experimenting the domain specific user preferences to reduce the false negatives in the user inbox it is worth-while to understand the background, training model and learning methods of Spam Bayes. Spam Bayes architecture works on couple of parts; firstly it does tokenisation where it takes an email message and breaks it up into tokens or words or features and secondly it then it does the scoring and calculating the combined score for the message. Finally it compares the combined score against a threshold to classify a message.

SpamBayes counts occurrences of tokens in spam and non-spam emails and learns which tokens are more indicative of each class. To predict whether a new email is spam or not, SpamBayes uses a statistical test to determine whether the email's tokens are sufficiently indicative of one class or the other, and returns its decision or unsure. In this section, we detail the statistical method SpamBayes uses to learn token scores and combine them in prediction, but first we discuss realistic models for deploying SpamBayes.

SpamBayes has a training model that works like this: a training set of labelled messages as Spam or Ham is fed into Spam Bayes and it then produces a classifier from those examples. This classifier (or filter) is subsequently used to classify future email messages that are received as spam or ham. Hence Spam Bayes labels messages after it classifies them, the messages that are clearly classified as spam are labelled as Spam, ones that are clearly classified as good are labelled as Ham and routed to Inbox, SpamBayes also has a third label—when it isn't confident one way or the other, it returns unsure and label then as Junk suspects. We adopt this terminology: the true class of an email can be ham or spam, and a classifier produces the labels ham, spam, and unsure (junk suspects).

There are three natural choices for how to treat unsure-labelled messages: they can be placed in the spam folder, they can be left in the user's inbox, or they can be put into a third folder called Junk Suspects for separate review by the user. The user can then go through this folder of unsure messages to either mark them as spam or ham. Sometimes classifying them as spam or ham can confuse the classifier as those messages contains the mixed features of spam and ham. Hence, another choice is to leave them as it is in unsure folder, the purpose here is not to contaminate the training of the filter with these messages.

4.3 Spam Bayes Learning Method

SpamBayes is a content-based spam filter that classifies messages based on the tokens (including header tokens) observed in an email. Based on ideas by Graham(Graham, 2002), Robinson(Robinson, 2003) developed the spam classification model together with Fisher's method for combining independent significance tests which is used by SpamBayes. Intuitively, SpamBayes learns how strongly each token indicates ham or spam by counting the number of each type of email that token appears in. When classifying a new email, SpamBayes looks at all of its tokens and uses a statistical test to decide whether they indicate one label or the other with sufficient confidence; if not, SpamBayes returns unsure.

SpamBayes tokenizes each email E based on words, URL components, header elements, and other character sequences that appear in E. Each is treated as a unique token of the email. The SpamBayes algorithm only records whether or not a token occurs in the message, not how many times it occurs.

Email E is represented as a binary vector e where

$$e_i = \begin{cases} 1 & \text{the } i^{th} \text{ token occurs in E} \\ 0 & \text{otherwise} \end{cases}$$

Below, we use e to refer to both the original message E and SpamBayes' representation of it since we are only concerned with the latter.

In training, SpamBayes computes a spam score vector P(S) where the i^{th} component is a token spam score for the i^{th} token given by

$$P_{(s,i)} = \frac{N_H \, N_S\,(i)}{N_H \, N_S\,(i) + N_S \, N_H\,(i)} \qquad (1)$$

where N_S, N_H, $N_{S(i)}$, and $N_{H(i)}$ are the number of spam emails, ham emails, spam emails including the i^{th} token and ham emails including the i^{th} token, respectively. The quantity $P_{(S,i)}$ is an estimate of Pr(E is spam | e_i) if the prior of ham and spam are equal, but for our purposes, it is simply a per-token score for the email. An analogous token ham score is given by $P_{(H,i)}= 1 - P_{(S,i)}$.

$P_{(S,i)}$ is smoothed through a convex combination with a prior belief x (default value of x = 0.5), weighting the quantities by $N_{(i)}$ (the number of training emails with the i^{th} token) and s (chosen for strength of prior with a default of s = 1), respectively(Robinson, 2003):

$$f_i = \frac{sx + N(i)}{s + N(i)} \quad P_{(S,i)} \qquad (2)$$

Effectively, smoothing reduces the impact that low token counts have on the scores. For rare tokens, the score is dominated by the prior x. However, as more tokens are observed, the smoothed score gradually shifts to the score in Eq. (1). An analogous smoothed ham score is given by 1 - f.

After training, the filter computes the overall spam score S(m) of a new message M using Fisher's method [7] for combining the scores of the tokens observed in M. SpamBayes uses at most 150 tokens from M with scores furthest from 0.5 and outside the interval [0.4,0.6]. Let δ(m) be the binary vector where δ(m)i = 1 if token i is one of these tokens, and 0 otherwise. The token spam scores are combined into a message spam score for M by

$$S(m) = 1- \chi^2_{2n}\,(-2(\log f)^T\, \delta\,(m))$$

where n is the number of tokens in M and $\chi^2_{2n}(\bullet)$ denotes the cumulative distribution function of the chi-square distribution with 2n degrees of freedom.

A ham score H(e) is similarly defined by replacing f with 1 - f in Eq. (3). Finally, SpamBayes constructs an overall spam score for M by averaging S(m) and 1 - H(m) (both being indicators of whether m is spam) giving the final score

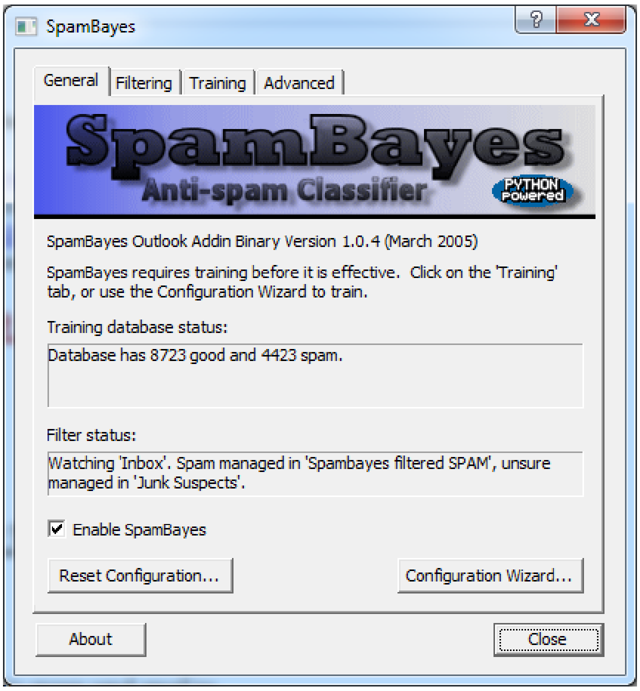
$$I(m) = \frac{1+ S(m) - H(m)}{2} \qquad (4)$$

for a message; a quantity between 0 (ham) and 1 (spam). SpamBayes predicts by thresholding I(m) against two user-tunable thresholds θ_0 and θ_1, with defaults $\theta_0 = 0.15$ and $\theta_1 = 0.9$. SpamBayes predicts ham, unsure, or spam if I(m) falls into the interval [0, θ 0], (θ_0, θ_1], or (θ_1,1], respectively, and filters the message accordingly.

5 Experimenting Spam Bayes

This experiment has been conducted with a purpose to test if the training of the Spam Bayes tool with domain specific user preferences takes effect or not. Spam Bayes was installed on Outlook client of the user in August 2012 and was trained with 8723 Ham and 4423 Spam data (Figure 1) that has been carefully collected for the purpose since March 2008. This training data is collected on the basis of individual user preference over a period of 5+ years belonging to a particular Domain.

The user has email filtering happening at the mail server which filter incoming email for multiple users but false negatives are escaping the server. The reason for this is that the features identified by filters are common to all the users and also that those filters have not been trained on the basis of domain specific keywords/features. In an aim to test if this idea will yield the results we installed SpamBayes at the users Outlook client. Spam Bayes will provide additional client based filtering on top of the existing content filtering done at the mail server. This means that Spam Bayes will attempt to filter false negatives that have escaped the content filters at the mail server and arriving users Inbox. This section is structured as Dataset, Experiment and Outcome.



SpamBayes Outlook Addin Binary Version 1.0.4 (March 2005)

SpamBayes requires training before it is effective. Click on the 'Training' tab, or use the Configuration Wizard to train.

Training database status:

Database has 8723 good and 4423 spam.

Filter status:

Watching 'Inbox'. Spam managed in 'Spambayes filtered SPAM', unsure managed in 'Junk Suspects'.

☑ Enable SpamBayes

Figure 1: Training in Aug 2012- initial dataset

5.1 Training Dataset

Spam has been coming to user's inboxes as email for a long time. In an attempt to catch the spam messages coming to the inboxes, it is very important that we use the appropriate samples. Since we are focusing on user belonging to an organisation belonging to a particular domain, let us refer to the business email counts worldwide. Given the amount of business emails sent per day, (Levenstein, 2013) (given in introduction section), if a spam corpus consists of 100,000 messages per day still only constitutes of 1/10 thousandth of the business email traffic globally. Hence the dataset that we use would be extrapolation of generalisation to be made in the spam dataset(Pitsillidis et al., 2012). It is very important to answer the question: is the available data that is being input to train the filer sufficient to reach a conclusion? Is the dataset too broad? Will it capture and feed the behaviour we want our filter to be trained for? Is the sample unbiased enough to capture the behaviour sincerely? Hence, to prepare a dataset that is not too large or too small. We would not address the issue of biased as we are trying to collect data specific to a domain. The focus here is to train the filter that it identifies the incoming email belonging to specific domain as ham and all those keywords and features that commonly exist in the spam data fed to the filter should be clearly identified as spam.

Given these considerations, in 2008, we started collected the false negatives that arrived in the inbox of the user. A separate folder called 'InboxJunk-Spam' was created and once identified as spam; the message was moved to that folder. Since the domain under consideration was 'Educational', careful attention was paid while classifying the unwanted messages arriving in the inbox as spam. This was done so that the filter does not get contaminated with those messages that belong to the domain. At the same time spam messages that were related to the education were classified as spam and moved to InboxJunk-Spam. Hence, the user does not keep getting spam email related to the domain. There were 1463 such messages identified, classified and moved to InboxJunk-Spam folder that was used for training purpose.

5.2 Experiment with Domain specific User Preference Training data

For the experimental purpose, we installed Spam Bayes on users Outlook client in August 2012. The training of the tool was done from Junk E-mail classified by the organisational filters and InboxJunk-Spam folders as Spam data and email in the Inbox and its sub folders at the Ham data. The initial training included 8723 Ham and 4423 Spam messages. Once the training is done, Spam Bayes trained with the domain specific user preference data will then filter all incoming email messages on the basis of this specific training. The tool will classify and categorise them as Spam or unsure or let it land in the User Inbox as (False negative). TO observe and record the results correctly few folders were created. We created sub folders for InboxJunk-Spam folder as SpamBayes filtered spam and InboxSpamsinceAug2012 so that the future messages arriving inbox as false negatives do not get mixed up in the same folder. The folders to observe the results of training are as:

InboxJunk-Spam – folder to collect all email that are Spam and gets through the organisational filters into the Inbox as legitimate (false negatives)

Spam Bayes Filtered Spam – incoming emails in the user inbox that were identified by Spam Bayes as Spam. These email collected are the email that are not identified by the mail server filters and arrive the users inbox as legitimate email(false negatives) but identified by Spam Bayes as Spam and moved to this folder

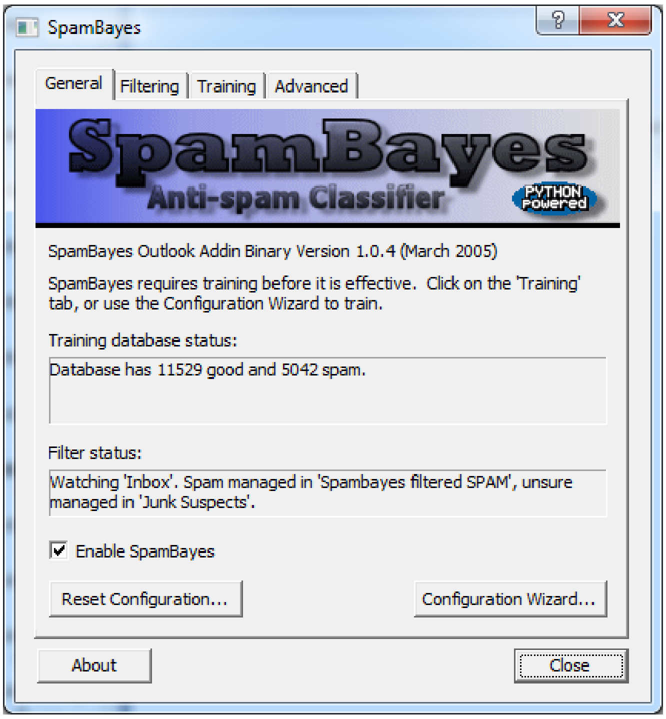
InboxSpamsinceAug2012- incoming emails that are false negatives, stay false negatives and land into users' inbox as legitimate emails. The trained tool could not identify these emails as Spam. Such emails were manually moved by the user to this folder.

Junk Suspects- Folder created by Spam Bayes tool to collect email that are suspected by the tool as Spam. User can go through this folder to report the email as Spam or Ham. This folder would contain False Positives too. This information is then used by the Spam Bayes tool to further train the filter.

It was then observed for a period of 12 months to identify if there was any improvement in the number of false positives arriving the user's Inbox. The following results were recorded.

The folder 'Spam Bayes filtered Spam' collected 683 emails that were going to Inbox as False negative. The folder InboxSpamsinceAug2012 collected 170 emails which was not identified by Spam Bayes as Spam and are still false negatives.

The folder Junk Suspects collected 408 emails that were otherwise going to Inbox as False negative. Out of these Junk Suspects user identified 10 emails as False Positives, these emails were not important email that would if missed would impact the productivity of the user. An important point to note is that once we trained the tool with the initial dataset, we did not retrain the filter till the end of observation period not to contaminate the learning of the filter with further data. The filter is now retrained with 11529 ham and 5042 spam message which is refined user specific data as shown in figure 2. Hence, Spam Bayes is now retrained with additional ham and spam messages collected in the last 12 months.



SpamBayes Outlook Addin Binary Version 1.0.4 (March 2005)

SpamBayes requires training before it is effective. Click on the 'Training' tab, or use the Configuration Wizard to train.

Training database status:

Database has 11529 good and 5042 spam.

Filter status:

Watching 'Inbox'. Spam managed in 'Spambayes filtered SPAM', unsure managed in 'Junk Suspects'.

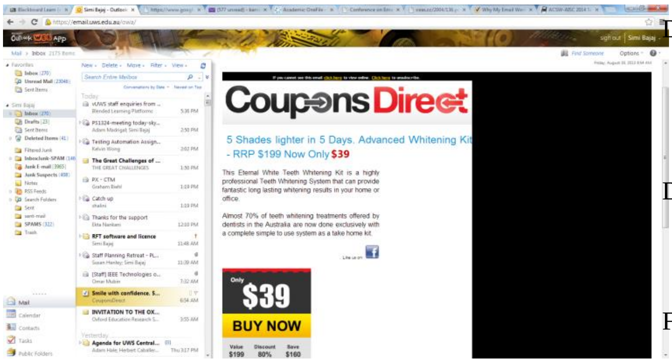
☑ Enable SpamBayes

Figure 2: Retraining after 12 months

5.3 Outcome

From this data, we can conclude that as a result of training the Spam Bayes filter with domain specific User preference data, the false negatives were reduced by 86% in the user inbox. This increases the user productivity and motivation by many hours.

Among the recommendations to be made in this paper is to re-emphasise the need to train the filters with user preference domain specific data so that the emails that pass through the filters as generic email can be caught. This makes the filters at the client level effective and efficient hence increasing the user productivity. Although this may not stop the false negatives completely as 14% of them still passed through the tool but it would deter some or help disabuse or reduce the inconvenience caused by spamming. An important point to be made here though, is that most of the false negatives in that 14% were related to educational domain only. Others did not have the features that do not common words such as offer, sale, buy (see figure); it is interesting to note the use of words such as buy in image to deceive the filters learning of such words.



If you cannot see this email click here to view online. Click here to unsubscribe.

CouponsDirect

5 Shades lighter in 5 Days. Advanced Whitening Kit - RRP $199 Now Only $39

This Eternal White Teeth Whitening Kit is a highly professional Teeth Whitening System that can provide fantastic long lasting whitening results in your home or office.

Almost 70% of teeth whitening treatments offered by dentists in the Australia are now done exclusively with a complete simple to use system as a take home kit.

Like us on

Only $39

BUY NOW

Value	Discount	Save
$199	80%	$160

Hence, we can conclude that training the tool with domain specific user preferred data did eliminate the false positives from the inbox that were unrelated (such as messages selling viagr@ or other pharmaceutical items, banks, holiday deals etc.) to the domain.

6 Conclusion and Future Work

Email spam has been the focus of studies for a long time. Though there are many different techniques to block spam email messages to reach users inbox, filtering is the most commonly used mechanism and has gained success to some extent. Given the large number of usage of email worldwide, email spam is still plentiful and scale of the problem is enormous. Researchers and organisations make the filers smart and self-learning but spammers are a step ahead. They keep on finding techniques to deceive the filters and their learning mechanisms. Hence, the problem still remains giving scope for researchers to work in the area. This work is an effort in the same scope to reduce false negatives/spam in the inbox of the users which has deceived the organisational filters. It is observed that this further filtering by training the filer with user specific data did make a difference in the amount of false positives.

Future work involves creating the feature sets including creating domain specific keywords and list of organisations which can be fed to the filter, conducting experiments and then observing the results to record the improvements.

7 References

BLANZIERI, E. & BRYL, A. 2008. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 29**,** 63-92.

CARUANA, G. & LI, M. 2012. A survey of emerging approaches to spam filtering. *ACM Comput. Surv.*, 44**,** 1-27.

CLIFFORD, M., FAIGIN, D., BISHOP, M. & BRUTCH, T. Miracle Cures and Toner Cartridges: Finding Solutions to the Spam Problem. 19th annual computer security applications conference (ACSAC 2003), 2003.

DEEPAK, P. & SANDEEP, P. Spam filtering using spam mail communities. *In:* SANDEEP, P., ed. Applications and the Internet, 2005. Proceedings. The 2005 Symposium on, 2005. 377-383.

DHINAKARAN, C., CHAE, C.-J. & LEE, J.-K. An Empirical Study of Spam and Spam Vulnerable email Accounts. *In:* CHAE, C.-J., ed. Future generation communication and networking (fgcn 2007), 2007. 408-413.

DU TOIT, T. & KRUGER, H. Filtering spam e-mail with Generalized Additive Neural Networks. Information Security for South Africa (ISSA), 2012, 15-17 Aug. 2012 2012. 1-8.

FERRIS_RESEARCH. 2007. Spam Control: The Current Landscape. Available: http://www.ferris.com/2007/01/02/the-commodity-status-of-spam-control/.

FORD, R. & SPATTORD, E. H. 2007. Happy Birthday, Dear Viruses. *Science*, 317**,** 210-211.

GARTNER. 2009. Gartner Says Number of Phishing Attacks on U.S. Consumers Increased 40 Percent in 2008 , 14 April 2009. *2009 Press Release* [Online]. [Accessed 23 Feb 2011].

GINA. 2010. Spam statistics of third-quarter 2010. *Panda Security Report* [Online]. Available: http://www.spywared.com/news/spam-statistics-of-third-quarter-2010-728.html.

GOMEZ, J. & MOENS, M.-F. 2010. Using Biased Discriminant Analysis for Email Filtering. *In:* SETCHI, R., JORDANOV, I., HOWLETT, R. & JAIN, L. (eds.) *Knowledge-Based and Intelligent Information and Engineering Systems.* Springer Berlin Heidelberg.

GRAHAM, P. 2002. A Plan for Spam. Available: http://www.paulgraham.com/spam.html.

GUZELLA, T. S. & CAMINHAS, W. M. 2009. A review of machine learning approaches to Spam filtering. *Expert Systems with Applications*, 36**,** 10206-10222.

HORIE, M. & NEVILLE, S. W. 2008. Addressing Spam at the Systems-level through a Peered Overlay

Network-Based Approach. *Novel Algorithms and Techniques In Telecommunications, Automation and Industrial Electronics.*

HUAI-BIN, W., YING, Y. & ZHEN, L. 2005. SVM Classifier Incorporating Feature Selection Using GA for Spam Detection. *Embedded and Ubiquitous Computing.*

ISLAM, M. & ZHOU, W. 2007. Architecture of Adaptive Spam Filtering Based on Machine Learning Algorithms

Algorithms and Architectures for Parallel Processing. *In:* JIN, H., RANA, O., PAN, Y. & PRASANNA, V. (eds.). Springer Berlin / Heidelberg.

J., W. & T., D. 2008. Research in Anti-Spam Method Based on Bayesian Filtering. *Pacific-Asia Workshop on Computational Intelligence and Industrial Application (PACIIA '08)*

JAEYEON, J. & EMIL, S. 2004. An empirical study of spam traffic and the use of DNS black lists. *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement.* Taormina, Sicily, Italy: ACM.

JEFF. 2013. Five Predictions Regarding Spam in 2013. *Articles* [Online]. Available: http://www.allspammedup.com/2013/01/five-predictions-regarding-spam-in-2013/ [Accessed 20 August 2013].

JUNG, J. & EMIL, S. An Empirical Study of Spam Traffic and the Use of DNS Black Lists. Internet Measurement Conference,, October 2004 Taormina, Italy.

KIM, H.-J., SHRESTHA, J., KIM, H.-N. & JO, G.-S. 2006. User Action Based Adaptive Learning with Weighted Bayesian Classification for Filtering Spam Mail. *In:* SATTAR, A. & KANG, B.-H. (eds.) *AI 2006: Advances in Artificial Intelligence.* Springer Berlin Heidelberg.

KIM, J., DOU, D., LIU, H. & KWAK, D. 2007. Constructing a User Preference Ontology for Anti-spam Mail Systems. *In:* KOBTI, Z. & WU, D. (eds.) *Advances in Artificial Intelligence.* Springer Berlin Heidelberg.

KLONOWSKI, M. & STRUMIŃSKI, T. 2008. Proofs of Communication and Its Application for Fighting Spam. *SOFSEM 2008: Theory and Practice of Computer Science.*

LAI, C.-C. 2007. An empirical study of three machine learning methods for spam filtering. *Knowledge-Based Systems,* 20**,** 249-254.

LEAVITT, N. 2007. Vendors Fight Spam's Sudden Rise. *Computer,* 40**,** 16-19.

LEVENSTEIN, J. 2013. Email statistics report, 2013-2017. *In:* RADICATI, S. (ed.) *Reports.* 1900 EMBARCADERO ROAD, SUITE 206. • PALO ALTO, CA 94303: Radicate Group Inc.

LIANG, T. & YU, Q. Spam Feature Selection Based on the Improved Mutual Information Algorithm. Multimedia Information Networking and Security (MINES), 2012 Fourth International Conference on, 2-4 Nov. 2012 2012. 67-70.

LIU, P., DONG, J.-S. & ZHAO, W. 2007. A Statistical Spam Filtering Scheme Based on Grid Platform. *Theoretical Advances and Applications of Fuzzy Logic and Soft Computing.*

MCGIBNEY, J. & BOTVICH, D. 2007. Establishing Trust Between Mail Servers to Improve Spam Filtering. *Autonomic and Trusted Computing.*

MEYER , T. A. & WHATELEY, B. SpamBayes: Effective open-source, Bayesian based, email classification system. . *In:* RESEARCH), D. H. M., ed. First Conference on Email and Anti-Spam (CEAS) 2004 Mountain View, CA

MOON, J., SHON, T., SEO, J., KIM, J. & SEO, J. 2004. An Approach for Spam E-mail Detection with Support Vector Machine and n-Gram Indexing. *Computer and Information Sciences - ISCIS 2004.*

NAGAMALAI, D., DHINAKARAN, C. & LEE, J. K. Multi Layer Approach to Defend DDoS Attacks Caused by Spam. *In:* DHINAKARAN, C., ed. Multimedia and Ubiquitous Engineering, 2007. MUE '07. International Conference on, 2007. 97-102.

NAKULAS, A., EKONOMOU, L., KOURTESI, S., FOTIS, G. & ZOULIAS, E. 2009. A Review of Techniques to Counter Spam and Spit. *In:* MASTORAKIS, N., MLADENOV, V. & KONTARGYRI, V. T. (eds.) *Proceedings of the European Computing Conference.* Springer US.

NAMESTNIKOVA, M. 2011. Spam in May 2011. *Analysis* [Online]. Available: http://www.securelist.com/en/analysis/204792179/Spam_in_May_2011 [Accessed 7 July 2011].

NAZIROVA, S. 2011. Survey on spam filtering techniques. *Communications and Network,* 3**,** 153+.

NELSON, B., BARRENO, M., JACK CHI, F., JOSEPH, A. D., RUBINSTEIN, B. I. P., SAINI, U., SUTTON, C., TYGAR, J. D. & XIA, K. 2009. Misleading Learners: Co-opting Your Spam Filter. *Machine Learning in Cyber Trust.*

NHUNG, N. & PHUONG, T. 2007. An Efficient Method for Filtering Image-Based Spam E-mail. *Computer Analysis of Images and Patterns.*

PASWAN, M. K., BALA, P. S. & AGHILA, G. Spam filtering: Comparative analysis of filtering techniques. Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on, 30-31 March 2012 2012. 170-176.

PITSILLIDIS, A., KANICH, C., VOELKER, G. M., LEVCHENKO, K. & SAVAGE, S. 2012. Taster's choice: a comparative analysis of spam feeds. *Proceedings of the 2012 ACM conference on Internet measurement conference.* Boston, Massachusetts, USA: ACM.

RAJENDRAN, B. & PANDEY, A. K. 2012. Contextual Strategies for Detecting Spam in Academic Portals

Advances in Computer Science and Information Technology. Computer Science and Engineering. *In:* MEGHANATHAN, N., CHAKI, N. & NAGAMALAI, D. (eds.). Springer Berlin Heidelberg.

ROBINSON, G. 2003. A Statistical Approach to the Spam Problem. Available: http://www.linuxjournal.com/article/6467 [Accessed 15 August 2013].

SIPONEN, M. & STUCKE, C. 2006. Effective Anti Spam Strategies in Companies: An International Study. 39th Hawaiia International Conference on System Sciences, 2006. IEEE.

SO YOUNG, P. & SHIN GAK, K. Labeling System for Countering SIP spam. Advanced Communication Technology, 2008. ICACT 2008. 10th International Conference on, 17-20 Feb. 2008 2008. 1644-1646.

SPIDERLABS. 2013. *Spam Types* [Online]. Trust Wave. Available: https://www.trustwave.com/support/labs/spam_types.asp [Accessed 6 July 2013].

TAKEMURA, T. & EBARA, H. 2008. Spam Mail Reduces Economic Effects. *Second International Conference on the Digital Society.* Sainte Luce, Martinique IEEE Computer Society Press.

TAKUMI, I., AKIRA, H., YOSHIAKI, K., ICHIMURA, T. A. I. T., HARA, A. A. H. A. & KUROSAWA, Y. A. K. Y. A classification method for spam e-mail by Self-Organizing Map and automatically defined groups. *In:* AKIRA, H., ed. Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on, 2007. 2044-2049.

VU DUC, L. & TRUONG NGUYEN, V. Bayesian spam filtering for Vietnamese emails. Computer & Information Science (ICCIS), 2012 International Conference on, 12-14 June 2012 2012. 190-193.

WEI, Z., FENG, G., DI, L. & FENG, X. Active learning based spam filtering method. Intelligent Control and Automation (WCICA), 2010 8th World Congress on, 7-9 July 2010 2010. 3302-3306.

WU, C.-H. & TSAI, C.-H. Robust classification for spam filtering by back-propagation neural networks using behavior-based features. *Applied Intelligence*.

XIAO-WEI, W. & ZHONG-FENG, W. Good word attack spam filtering model based on artificial immune system. Automatic Control and Artificial Intelligence (ACAI 2012), International Conference on, 3-5 March 2012 2012. 1106-1109.

XIAO, L., JUNYONG, L. & MEIJUAN, Y. E-Mail Filtering Based on Analysis of Structural Features and Text Classification. Intelligent Systems and Applications (ISA), 2010 2nd International Workshop on, 22-23 May 2010 2010. 1-4.

YU, B. & XU, Z.-B. 2008. A comparative study for content-based dynamic spam classification using four machine learning algorithms. *Knowledge-Based Systems,* 21**,** 355-362.

ZHANG, Y., YANG, X. & LIU, Y. Improvement and optimization of spam text filtering system. Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on, 29-31 Dec. 2012 2012. 448-451.