A survey of classifier techniques for Spamfiltering

Dr. VivekKapoor
Asst.Professor, Dept.Of Information Technology,
Institute Of Engineering & Technology ,
Devi Ahilya University, Indore, India.

Rahul Maheshwari
Dept. Of Computer Engineering ,
Institute Of Engineering & Technology,
Devi Ahilya University, Indore, India.

***Summary*–Email spam or junk e-mail (unwanted e-mail "usually of a commercial nature sent out in bulk") is one of the majorproblems of the today's Internet, bringing financial damage to companies and annoying individual users. Among theapproaches developed to stop spam, filtering is an important and popular one. Common uses for mail filters include organizing incoming email and removal of spam and computer viruses. A less common use is to inspect outgoing email at some companies to ensure that employees comply with appropriate laws. Users might also employ a mail filter to prioritize messages, and to sort them into folders based on subject matter or other criteria. Mailfilters can be installed by the user, either as separate programs, or as part of their email program (email client). In email programs, users can make personal, "manual" filters that then automatically filter mail according to the chosen criteria. In this paper, we present a survey of the performance of five commonly used machine learning methods in spam filtering. Most email programs now also have an automatic spam filtering function.**

Key words- E-mail classification, Spam, Spam filtering, Machine learning,algorithms

.

1. INTRODUCTION

Email was widely accepted by the business community as the first broad electronic communication medium and was the first 'e-revolution' in business communication. Email is very simple to understand and like postal mail, email solves two basic problems of communication: logistics and synchronization. **Electronic mail**, most commonly referred to as **email** or **e-mail** since c. 1993, is a method of exchanging digital messages from an author to one or more recipients. Email is a short word for electronic mail. You **create** texts and send them over a network of computers. The first emails go back to the 1960s. The **invention** has **influenced** our lives and emails have become a **popular** means of communication.An Internet email message consists of three components, message *envelope*, message *header*, and message *body*.It uses technology to communicate a digital message over the Internet. Users use email differently, based on how they think about it.

```
From zsuthiongie@invitation.sms.ac Fri Mar 11
18:02:00 2005
Return-Path: <zsuthiongie@invitation.sms.ac>
Received: from smtp57.sms.ac (local host [127.0.0.1])
by mail.nutn.edu.tw (8.12.10+Sun/8.12.9) with ESMTP
id
j2BA1v5t010627
for <cclai@mail.nutn.edu.tw>; Fri, 11 Mar 2005
18:01:59 +0800
(CST)
X-Authentication-Warning: mail.nutn.edu.tw: is can
owned process
doing -bs
Received: from LOCALHOST (unknown [10.1.4.231])
by smtp57.sms.ac (Postfix) with SMTP id 01EFE3825B
for <cclai@mail.nutn.edu.tw>; Fri, 11 Mar 2005
05:00:47 -0500
(EST)
SUBJECT: zsuthiongie(3rd request)
To: cclai@mail.nutn.edu.tw
CONTENT-TYPE: text/plain
Message-Id:
<20050311100047.01EFE3825B@smtp57.sms.ac>
Date: Fri, 11 Mar 2005 05:00:47 -0500 (EST)
From: zsuthiongie@invitation.sms.ac
Content-Length: 441
Status: R
```

Fig.-an example of the header in an e-mail.

1.1 Structure Of An E-Mail

In addition to the body message of an e-mail, an e-mail has another part called the header. The job of the headers to store information about the message and it contains many fields, for example, tracing information aboutwhich a message has passed:

1.1.1 Received- authors or persons taking responsibilityfor the message

1.1.2 From - intending to show the envelop address of the real sender as opposed to the sender used for replying

ISSN: 2278 – 1323

1.1.3 Return-Path - unique of ID of this message

1.1.4 Message-ID - format of content

1.1.5 Content-Type - format of content

There are 3 different types of emails, this table summarizes the situation.

	Marketing = Bulk	*Notification = Trigger/ Auto-responder/Alerts*	*Transactional*
Trigger	Sender	Sender or Recipient *Event scenario/system*	Recipient
Relation	One-to-many	One-to-one	One-to-one
Unsubscribe link	Yes	Yes	No

1.1.1 Marketing Emails

Marketing (or Bulk) emails stimulate your clients and leads. They contain informative / incentive messages. The recipient must agree to receive such emails: opt-in is mandatory.However, the recipient does not make an explicit request for a message in particular. For example: he doesn't subscribe for the "November Newsletter", he rather subscribes to the "Monthly Newsletter".Common examples of **marketing emails**:NewslettersFlashsales,Sales/promotions announcements.

1.1.2. Notification Emails

Notification email is also known as trigger, alert or auto-responder. They allow the user to be notified each time a particular event happens (or has happened). More generally, the notification email may be used in order to celebrate and/or mark an event. From a marketer's point of view, it can be relevant to encourage the targets to opt in to receive notifications about the services being offered. Think of an email such as "Mr. X is now following you on Twitter". This kind of message is more often opened and it motivates the recipient into checking their account. Common examples of **notification emails**: Getting in touch a few days after registration, Congratulations after a status change (first purchase, subscription...), Birthday email

1.1.3. Transactional Emails

This is an expected message and its content is information that the client wishes to check or confirm, and not "discover". This type of email is not intended to optimize the customer relationship but to define it and mark it out. It is a point of reference in one's CRM. Common examples of **transactional emails**: Welcome message / Account opening, Shipment tracking and order status, Order shipment confirmation

2. WHAT IS SPAM MAIL

The term spam refers to submitting the same message to a large group of individuals in an effort to force .Spam simply means that people are sending emails to people that haven't requested them.A study estimated that over 70% of today's business emails are spam [Aladdin Knowledge Systems,2011]; If you suffer from spam then you might be looking for a way to stop the spam so that you can carry on with more important tasks.

There are two types of spam:

(a) Cancelable Usenetspam refers to spam email in which a single message is sent to 2 or more usenet groups. This type of spam is directed at “lurkers”, or individuals who read newsgroups but who either do not or infrequently post or give their email addresses away. Cancelable Usenet spam reduces the utility of newsgroups by forcing through advertising, and as such decreases the ability of administrators and managers of newsgroups to manage accepted topics. This spam is run at a low cost to those sending out spam.

(b)Email spam refers to spam email that is directed at individual users with direct addresses; email spam lists are usually created by scanning Usenet postings, sterling Internet mailing lists or searching the Web for addresses. A variant of this form of spam is sent directly to mailing lists and email discussions that are used by public and private forums. Email spam costs individuals submitting spam email money; for example, ISPs and online services need to pay to transmit spam directly to subscribers.

In addition, there are three main components to all types of spam:

Anonymity-the sender's identity and address are concealed

Mass mailing-spam email is sent to a large number of recipients and in high quantities

Unsolicited-the individuals receiving spam would otherwise not have opted to receive it

Common forms of spam include commercial advertising, usually for dubious products, such as get-rich-quick schemes, quasi-legal services, political messages, chain letters and fake spam used to spread viruses. Spam mails vary significantly in content and they roughly belong to the following categories: money making scams, fat loss, improve business, sexually explicit, make friends, service provider advertisement, etc.[F. Smadja, H. Tumblin, 2002], One example of aspam mail is shown as in the Fig.

Date: Mon, 12 Sep. 2014 14:16:44 -0500
From: Ramadan Faraj<Ramadan_faraj@yahoo.com>
Subject: Those young people taking the position you deserve because you lack a
Degree?
To: XXX <xxx@yahoo.com>
Content-Type: text/plain; charset=iso-8859-1

WHAT A GREAT IDEA!
Ring anytime 1-404-549-4731
We provide a concept that will allow anyone with sufficient work experience toobtain a fully verifiable University Degree. Bachelors, Masters or even a Doctorate.
Think of it, within four to six weeks, you too could be a college graduate. Many
people share the same frustration, they are doing the work of the person that has the
Degree and the person that has the degree is getting all the money.
Don't you think that it is time you were paid fair compensation for the level of work
You are already doing?
This is your chance to finally make the right move and receive your due benefits.
If you are more than qualified with your experience, but are lacking that prestigious
piece of paper known as a diploma that is often the passport to success.
CALL US TODAY AND GIVE YOUR WORK EXPERIENCE THE CHANCE TO EARN YOU.

2.1 Cause of arrival

The usefulness of email is being threatened by four phenomena: email bombardment, spamming, phishing, and email worms.Spamming is unsolicited commercial (or bulk) email. Hundreds of active spammers sending this volume of mail results in information overload for many computer users who receive voluminous unsolicited email each day.

2.2The Costs of Spam: How Spam Affects Your Bottom Line(Negative effects)

The biggest cost caused by spam will be paid by the ISP's who have to pay increased bandwidth charges as a result of increase network traffic. However, spam also causes problems for many other people because of increased fraud, wasted time, and various other scams.

2.3Cost of Spam

Many people get annoyed at spam but they don't realize that it's also actually costing. We all have to waste a considerable amount of time manually sifting through our mailboxes to sort out what is genuine and what is spam. This time could be spent better by doing something else. Businesses lose billions of dollars as a result of spam. Spam blockers can be used to save everyone time; however these also have their own cost. Not every message that they flag as spam is actually spam. This can mean that it's very easy to miss important emails accidentally.

2.3.1 Decline in Productivity

Spam not only robs money from business owners, but also wastes the time of employees who were once productive. The average worker can spend up to 20 seconds evaluating and deleting spam messages one by one. This is typically the case in a work environment where spam is quarantined and left for the recipient to review at a later time. Even though the messages are being deleted, it still costs the business valuable time as employees attempt to determine the legitimacy of these emails.

2.3.2 Wasted Storage Space

Several businesses quarantine spam in order to get a handle on the problem. The truth is that the entire process requires an additional storage capacity in order to accommodate the email in question. For many online business owners, this is the best way to prevent spam from directly reaching their employees. Unfortunately, several users fail to even bother with quarantined messages, believing that they must have been isolated for a reason. The result:

extra storage space purchased for email to just sit until it is automatically deleted.

2.3.3 Costs for the Internet Service Provider

It is difficult to calculate how spam financially affects a service provider. Symantec Corp., a prominent vendor of anti-spam products, released a report in October of 2014, stating that an estimated 70% of all email we receive is spam. This very burden of email traffic is what has forced internet service providers to supply extra capacity to their network and servers.

.

3. BACKGROUND

3.1 What is spam filter?

A spam filter is a software program that blocks unwanted messages in 3 main ways -

(a)Establishing white and black lists-internet spam filter programs create an approved list (white list), where messages sent by approved users are sent, as well as a black list, where addresses not approved are stored.

(b)Blocking "sporm"- spam filters block a high percentage of spam pornography, as well as incoming mail that is deemed to be adult in content or contain adult images

(c)Organizing email-most filter programs let individuals create folders so that they can store different categories of emails accordingly (i.e. financial, personal, games). Incoming mail is automatically organized into the appropriate folder so that the user may choose what to read.

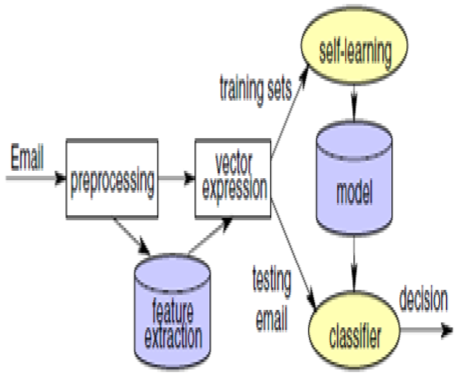


Fig.- The process of spam filtering

3.2 Anti-spam Techniques

Junk email can clog your inbox and waste your time there are several things that can do to reduce the amount of spam that makes it into our email inbox.

3.2.1 Filtering by Content

One way to reduce a significant amount of spam email is to filter it by word. Many spam emails are about certain products such as medicine meant to increase a man's sexual abilities. You can filter emails with those words or emails of a certain topic which you know you'll never or rarely receive legitimate information about.

3.2.2 Blacklisting

You can filter the messages you receive to blacklist certain IP addresses of common spammers or Internet hosts that shouldn't be sending external emails.

3.2.3 Avoid Email Harvesting

Do not put your true email address on lists and forms if you don't have to. This reduces the chance of spammers accessing your email and buying it from other spammers in a process called email harvesting. If you have your email address on your personal, professional website, online corporate directory or something similar, do not post the entire email in the form that it's used to send the message. Doing this reduces the chance of spam bots finding your email address online.

4. LITERATURE SURVEY

This section gives a brief overview of the underlying theory and implementations of the algorithms we consider. We shall discuss the Naïve Bayesian classifier, the k-NN classifier, the neural network classifier and the support Vector machine classifier.

4.1 Support Vector Machine

Support vector machines (SVMs) are relatively new techniques that have rapidly gained popularity because ofthe excellent results they have achieved in a wide variety of machine learning problems, and because they havesolidtheoretical underpinnings in statistical learning theory [N. Cristianini, B. Schoelkopf,2002].

Support vector machine (SVM) algorithms divide the dimensional space representation of the data into two regions using a hyper plane. This hyper plane always maximizes the margin between the two regions or classes. The margin is defined by the longest distance between the examples of the two classes and is computed based on the distance between the closest instances of both classes to the margin, which are called supporting vectors [V.Vapnik, 1998].Instead of using linear hyper planes, many implementations of these algorithms use so-called kernel functions. Thesekernel functions lead to non-linear classification surfaces, such as polynomial, radial or sigmoid surfaces [S. Amari, S. Wu, 1999].*Formal definition*- More formally, a support vector machine constructs a hyper plane or set of hyper planes in ahigh- or infinite- dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of anyclass (so-called functional margin), since in general the larger the margin the lower the generalization error of theclassifier.

4.2 Naïve Bayes Classifier

The Naive Bayes classifier is a simple statistical algorithm with a long history of providing surprisingly accurate results. It has been used in several spam classification studies [**I. Androutsopoulos, J. Koutsias,** 2000], and has become somewhat of benchmark. It gets its name from being based on Bayes' rule of conditional probability, combined with the “naive “assumption that all conditional probabilities are independent [**I. Witten, E. Frank,**2000].

Naive Bayes classifier examines all of the instance vectors from both classes. It calculates the prior class probabilities
as the proportion of all instances that are spam (Pr[spam]),and not-spam (Pr[not spam]). Then (assuming binary attributes) it estimates four conditional probabilities for each attribute: Pr[true|spam], Pr[false spam],Pr[true|notspam], and Pr[false|notspam]. These estimates are calculated based on the proportion of instances of the matching class that have the matching value for that attribute.

To classify an instance of unknown class, the “naive” version of Bayes’ rule is used to estimate first the probability of the instance belonging to the spam class, and then the probability of it belonging to the not-spam class. Then it normalizes the first to the sum of both to produce a spam confidence score between 0.0 and 1.0.Note that the denominator of Bayes’ rule can be omitted because it is cancelled out in the normalization step. Inters of implementation, the numerator tends to get quite small as the number of attributes grows, because so many tiny probabilities are being multiplied with each other. This can become a problem for finite precision floating-point numbers. The solution is to convert all probabilities to logs, and perform addition instead of multiplication. Note also that conditional probabilities of zero must be avoided; instead a “Laplace estimator” (a very small probability) is used.

It is important to note that using binary attributes in the instance vectors makes this algorithm both simpler andmore efficient. Also, given the prevalence of sparse instance vectors in text classification problems like thisone, binary attributes offer the opportunity to implement very significant performance optimizations. Fig. presentsthe Naive Bayes training and classification algorithms used.

```
Naive Bayes Training Algorithm:

priorProbSpam = proportion of training set that is spam
priorProbNotSpam = proportion of training set that is notspam
For each attribute i:
probTrueSpam[i] = prop. of spams with attribute i true
probFalseSpam[i] = prop. of spams with attribute i false
probTrueNotSpam[i] = prop. of not-spams with attribute i true
probFalseNotSpam[i] = prop. of not-spams with attribute i false
Naive Bayes Classification Algorithm:
probSpam = priorProbSpam
probNotSpam = priorProbNotSpam
For each attribute i:
if value of attribute i for message to be classified is true:
probSpam = probSpam × probTrueSpam[i]
probNotSpam = probNotSpam × probTrueNotSpam[i]
else:
probSpam = probSpam × probFalseSpam[i]
probNotSpam = probNotSpam × probFalseNotSpam[i]
spamminess = probSpam/(probSpam + probNotSpam)
```

Fig-: Naive Bayes training and classification algorithms.

4.3 Artificial Neural Networks

An artificial neural network (ANN), usually called neuralnetwork (NN), is a mathematical model or computationalmodel that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases an ANN is anadaptivesystem that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data.

By definition, a "neural network" is a collection ofinterconnected nodes or neurons. The bestknownexample of one is the human brain, the mostcomplexand sophisticated neural network. Thanks to thiscranial-based neural network, we are able to make veryrapid and reliable decisions in fractions of a second. [**C. Miller**, 2011]

InputHidden Output

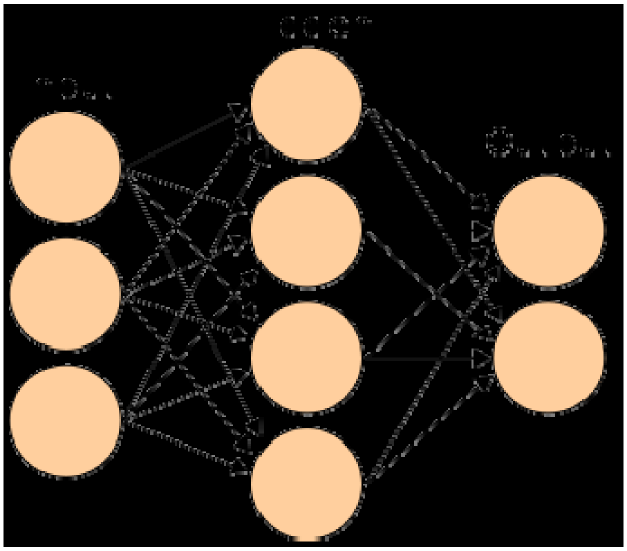


Fig.- an artificial neural network is an interconnected group of nodes, akin to the vast network of neurons in the human brain.

Spam presents a unique challenge for traditional filtering technologies: both in terms of the sheer number of messages (millions of messages daily) and in the breadth of content (from pornographic to products and services, to Finance). Add to that the fact that today's economic fabric depends on email communication – which is equally broad and plentiful and whose subject matter contextually overlaps with that of many spam messages – and we've got a serious challenge.

How it works- Since a neural network is based on pattern recognition, the underlying premise is that each message can be quantified according to a pattern. This is represented below in Fig. Each plot on the graph (alsoknown as a "vector") represents an email message. Although this 2-D example is an over-simplification, it helps to visualize the principle used behind neural networks.

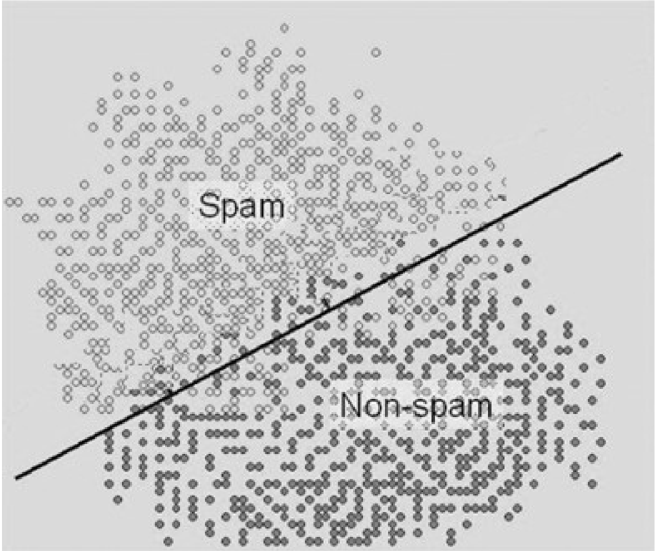


Fig.- Distinctive patterns of good and spam messages cluster into relatively distinct groups.

To identify these patterns, the neural network must first be "trained". This training involves a computational analysis of message content using large representative samples of both spam and non-spam messages. Essentially the network will "learn" to recognize what we humans mean by "spam" and "non-spam". To aid in this process, we first need to have a clear, concise definition of "spam":

Spam, n., email sent in bulk where there is no direct agreement in place between the recipient and the sender to receive email solicitation.

U.B.E. (Unsolicited Bulk Email) is another acronym for spam that effectively encapsulates this definition. To create training sets of spam and non-spam emails, each email is carefully reviewed according to this simple, yet restrictive definition of spam. Although the average user often considers all unwanted emails as "spam", emails that border on "solicited" (it was likely requested at some point by the user) should be rejected outright. Examples of these might include email sent from easily recognizable domains, such as Amazon.com or Yahoo.com. A good motto to follow here is: "when in doubt, throw it out". Similarly, non-spam email should be restricted to personal email communications between individuals or groups, and avoid any forms of bulk mailings, regardless of whether they were solicited or not. Once these sets have been gathered and approved, the neural network is ready for training.

The ANN system now preprocesses each email in the respective training sets to determine exactly which of these relevant words are found in each spam email, and which of these words are found in the non-spam email. Next, the ANN is trained to recognize certain combinations or patterns of interesting or relevant words to identify spam, or if it sees other combinations, to identify non-spam. The artificial neural network uses a set of sophisticated mathematical equations to perform this type of computation. As some spam and non-spam messages will often "share "characteristics, a clear distinction cannot always be made. By the "non-spam" plots or vectors that find themselves in the "spam" cluster and vice versa. In this "grey area" lies the potential for false positives. After the training is complete, the ANN can now be used to scan live-stream email. Each message is scanned to identify relevant words, which are then processed by the ANN. If the ANN again sees certain types of combinationsof word usage indicating a probability of spam, it will report spam, along with a probability value. Following theexample in Fig.., if the vector or plot computed for themessage landed above the dividing line, it would be considered "spam". Its probability or confidence level would depend on the relative distance away from the line. To maximize detections while avoiding false positives, well-designed heuristics engine will accommodate different sensitivity thresholds, or levels of aggressiveness, in identifying spam. What this means is that the cut-off or dividing point between spam and non-spam can be adjusted so that the likelihood of a false positive match will be greatly reduced. This can be seen in Fig. below.

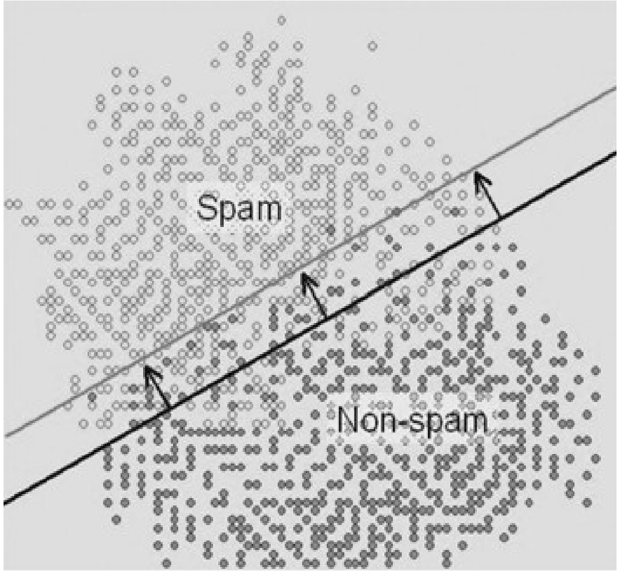


Fig- The sensitivity threshold can be adjusted to avoid the"grey" area.

In other words, the further away from the central dividing line between ham and spam email clusters, the lower the chance of false positive detections. Note in Fig. that there are far fewer non-spam vectors or patterns above the new cut-off or dividing line.

4.4 K-nearest neighbor classifier

The k-nearest neighbor (K-NN) classifier is considered an example-based classifier, that means that the training documents are used for comparison rather than an explicit category representation, such as the category profiles used by other classifiers. As such, there is no real training phase. When a new document needs to be categorized, the most similar documents (neighbors) are found and if a large enough proportion of them have been assigned to a certain category, the new document is also assigned to this category, otherwise not .

Additionally, finding the nearest neighbors can be quickened using traditional indexing methods. To decide whether a message is legitimate or not, we look at the class of the messages that are closest to it. The comparison between the vectors is a real time process. This is the idea of the k nearest neighbor algorithm:

Stage1. Training

Store the training messages.

Stage2. Filtering

Given a message x, determine its k nearest Neighbors among the messages in the trainingset. If there are more spam's among these neighbors, classify given message as spam.Otherwise classify it as legitimate mail.We should note that the use of an indexing method inorder to reduce the time of comparisons induces anupdate of the sample with a complexity O(m), where m isthe sample size. As all of the training examples are storedin memory, this technique is also referred to as a memorybasedclassifier [P. Cunningham, N. Nowlan,2011]. Another problem of the presented algorithm is that there seems to be no parameter that wecould tune to reduce the number of false positives. This problem is easily solved by changing the classification rule to the following l/k rule: If l or more messages among the k nearest neighbors of are spam, classify x as spam, otherwise classify it aslegitimatemail. The k nearest neighbor rule has found wide use in general classification tasks. It is also one of the few universally consistent classification rules.

5. REVIEW

We compare[d] and evaluated a range of classification techniques, with an assessment of their merits, disadvantages and range of application[s]."

5.1 Support Vector Machine

Advantages of SVMs -High accuracy, nice theoretical guarantees regarding over fitting, and with an appropriate kernel they can work well even if you're data isn't linearly separable in the base feature space.

Especially popular in text classification problems where very high-dimensional spaces are the norm. Memory-intensive and kind of annoying to run and tune, though, so I think random forests are starting to steal the crown.No distribution requirement, compute hinge loss, flexible selection of kernels for nonlinear correlation, not suffer multicollinearity, hard to interpretSupport Vector Machines work very well in many circumstances and perform very good with large amounts of data.

5.2 Naïve Bayes Classifier

Advantages of Naive Bayes - Super simple, you're just doing a bunch of counts. If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. And even if the NB assumption doesn't hold, a NB classifier still often performs surprisingly well in practice. A good bet if you want to do some kind of semi-supervised learning, or want something embarrassingly simple that performs pretty well.Naive Bayes mechanism is very simple to understand, it has also a high performance and is also easy to implement

5.3 Decision Tree Classifier

Advantages of Decision Trees - Easy to interpret and explain (for some people -- I'm not sure I fall into this camp). Non-parametric, so you don't have to worry about outliers or whether the data is linearly separable (e.g., decision trees easily take care of cases where you have class A at the low end of some feature x, class B in the mid-range of feature x, and A again at the high end). Their main disadvantage is that they easily over fit, but that's where ensemble methods like random forests (or boosted trees) come in. Plus, random forests are often the winner for lots of problems in classification (usually slightly ahead of SVMs, I believe), they're fast and scalable, and you don't have to worry about tuning a bunch of parameters like you do with SVMs, so they seem to be quite popular these days.No distribution requirement, heuristic, good for few categories variables, not suffer multicollinearity (by choosing one of them). Decision trees and rule based algorithms are good because you can understand the model that was built for classifying, unlike with neural networks.

5.4 Cons

(a) KNN should be avoided since the evaluation is quite heavy if your training dataset contains several thousand elements; although it gives really good results.

(b) Naive Bayes is very simple and quickly to evaluate but we had to tweak it to handle unbalanced classes. The Naive Bayes classifier is *simple, fast and of limited capability* when it comes to anything but the most*trivial*cases of classificationAll in all, I'd say that this is very data-dependent. Best technics depends on the data and what accuracy/efficiency trade off you is expecting.

(c)The *Support Vector Machine* is *complex, slow, takes a lot of memory*, but is an immensely *powerful* classifier. The *SVM-KA* also has the limitation of being a *binary* classifier.

All in all, we'd say that this is very data-dependent. Best technics depends on the data and what accuracy/efficiency trade off you is expecting.

How large is your training set?
If your training set is small, high bias/low variance classifiers (e.g., Naive Bayes) have an advantage over low bias/high variance classifiers (e.g., kNN or logistic regression), since the latter will over fit. But low bias/high variance classifiers start to win out as your training set grows (they have lower asymptotic error), since high bias classifiers aren't powerful enough to provide accurate models.

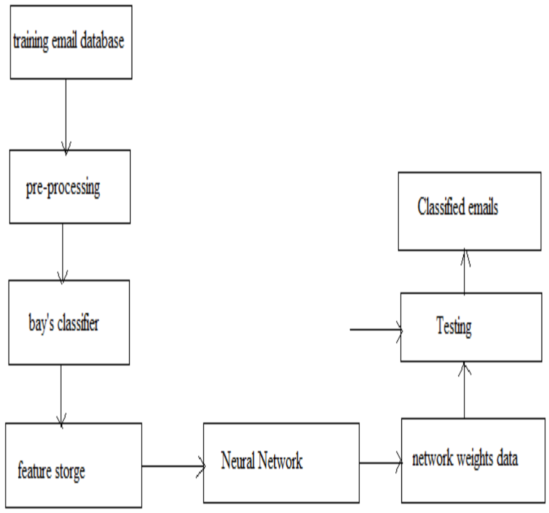
6. PROPOSED WORK

Now in these days the email and messaging is a routine work for most of the persons. Sometimes the mails come from the unauthenticated servers or malicious mails can harm the users privacy and security by phishing emails and their contents. Therefore an adoptive and secure technique is desired which can able to preserve the training feature for future uses. In addition of that system is able to detect the malicious emails with their spam filtering ability.

Many of the filtering techniques are basedon text categorization methods. Thus filtering spam turns on a classification problem. Roughly, we can distinguish between two methods of machine classification. The first one is done on some rules defined manually. This kind of classification can be used when all classes are static, and their components are easily separated according to some features.

The second one is done using machine learning techniques. It is more convenient when the characteristics of discrimination are not well defined. These techniques attempt to generate on a set of samples, quasi or semi automatically a classifier with an acceptable error rate.

The overview of the proposed systems components discussed in this section.



Training email database **-**the proposed system is a machine learning technique for classifying emails into four categories. Therefore in order to perform training of the data model a set of pre-labeled data is required for train the classification algorithms.

Pre-processing **-** the training data is cleaned and the undesired contents from the raw email data are removed. In this process the stop words and the frequent less weighted words are also eliminated from the email contents.

Bay's classifier - in this phase the bay's classifier is applied over the data by which two different probabilities is estimated for each words in data base. For example a word "ICICI bank" the probability to be in spam mail and the same words probability for become in a legitimate mail is estimated as features of the email training.

Feature storage **-**the extracted features from the bay's classifier is preserved for future use in neural network learning.

Neural network **-** that is a neural network learning phase where the neural network a word and their probabilities to be in spam mail and for legitimate mail.

Network weight data: after optimum training with all the words in the bay's database the neural network weights are preserved for future use. When new data set is added to the system the bay's classifiers weights are updated and the neural network weights are cleared and recomputed.

Testing - in this phase the system accepts the emails to classify, therefore the trained neural network load their trained weights and performed the classification.

Classified emails - in this phase the neural networks classification results are listed.

6.1 Product Perspective

There are two main prospective are for resolving in the proposed work.

6.1.1Identification of email contents or classifies the emails in terms of

6.1.2Mail is legitimate and contains normal URLs

6.1.3Mail is legitimate and contains malicious URL

6.1.4Mail is Spam and contains malicious URLs

6.1.5Mail is spam and contains normal URLs

7. CONCLUSION

Spam is becoming a very serious problem to the Internetcommunity, threatening both the integrity of the networks and the productivity of the users. In this paper, we proposefive machine learning methods for anti-spam filtering.In this paper we discussed the problem of spam and gave an overview of learning based spam filtering techniques.There is no common definition of what spam is, but most of the sources agree that the core feature of the phenomenon isthat spam messages are unsolicited. Spam causes a number of problems of both economical and ethical nature, whichresults in particular in the attempts of legislative definition and prohibition of spam. The most popular and well-developed approach to anti-spam is learning based filtering. The current state of threatIncludes many filters based on various classification techniques applied to different parts of email messages.

Email filtering is the processing of email to organize it according to specified criteria. Most often this refers to theautomatic processing of incoming messages, but the term also applies to the intervention of human intelligence in Addition to anti-spam techniques, and to outgoing emails as well as those being received.
Email filtering, software inputs email. For its output, it might pass the message through unchanged for delivery to the user's mailbox, redirects the message for delivery elsewhere, or even throws the message

away. Some mail filters are able to edit messages during processing.

In conclusion, we can say that the field of anti-spam protection is by now mature and well-developed. Then a question arises, why our inboxes are still often full of spam? Reactivity of spammers plays a role surely, and so does the complex nature of spam data. But one more issue not to be under estimated here is that we usually do not protect against spam in all the available ways. In other words, one point which should always be remembered by server administrators and end users is that the anti-spam technologies should be not only designed and developed, but also deployed and used.

9.References

[1] Aladdin Knowledge Systems, Anti-spam white paper,www.csisoft.com/security/aladdin/esafe_antispam_whitepaper.pdf Retrieved December 28, 2011.

[2] **F. Smadja, H. Tumblin,** "Automatic spam detection as a text classification task", in: Proc. of Workshop on Operational Text Classification Systems, 2002.

[3] **A. Hassanien, H. Al-Qaheri, "**Machine Learning in Spam Management", IEEE TRANS., VOL. X, NO. X, FEB.2009

[4] **P. Cunningham, N. Nowlan,** "A Case-Based Approach to Spam Filtering that Can Track Concept Drift", [Online]Available: https://www.cs.tcd.ie/publications/techreports/ reports.03/TCD-CS-2003-16.pdf RetrievedDecember 28, 2011

[5] **F. Roli, G. Fumera,** "The emerging role of visual pattern recognition in spam filtering: challenge and opportunity forIAPRresearchers"
http://www.iapr.org/members/newsletter/Newsletter07-02/index_files/Page465.htm Retrieved December 28, 2011

[6] **H. West**, "Getting it Wrong: Corporate America Spams the Afterlife". Clueless Mailers. (January 19, 2008).

[7] **B. Parizo**, "Image spam paints a troubling picture". Search Security. (2006-07-26)

[8] **Symantec** (2011) VBS.Davinia.B, [Online] Available: http://www.symantec.com/security_response/writeup.jsp?docid=2001-020713-3220-99 Retrieved December 28, 2011

[9] **I. Androutsopoulos, J. Koutsias,** "An evaluation of naïve bayesian anti-spam filtering". In Proceedings of the
Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning
(ECML 2000), pages 9–17, Barcelona, Spain, 2000.

[10] **I. Androutsopoulos, G. Paliouras,** "Learning to filter spamE-mail: A comparison of a naïve bayesian and a memorybasedapproach". In Proceedings of the Workshop on Machine Learning and Textual Information Access, 4thEuropean Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000), pages 1–
13, Lyon, France, 2000.

[11] **J. Hidalgo,** "Evaluating cost-sensitive unsolicited bulk email categorization". In Proceedings of SAC-02, 17th
ACM Symposium on Applied Computing, pages 615–620, Madrid, ES, 2002.

[12] **K. Schneider,** "A comparison of event models for naïve bayes anti-spam e-mail filtering". In Proceedings of the 10thConference of the European Chapter of the Association for Computational Linguistics, 2003.

[13] **I. Witten, E. Frank,** "Data Mining: Practical MachineLearning Tools and Techniques with Java Implementations".Morgan Kaufmann, 2000.

[14] **N. Cristianini, B. Schoelkopf,** "Support vector machines and kernel methods, the new generation of learning
machines". Artificial Intelligence Magazine, 23(3):31–41, 2002

[15] **V. Vapnik**, "The Nature of Statistical Learning Theory, Springer; 2 edition (December 14, 1998)

[16] **S. Amari, S. Wu**, "Improving support vector machine classifiers by modifying kernel functions". Neural Networks, 12, 783– 789. (1999).

[17] **C. Miller**, "Neural Network-based Antispam Heuristics" ,Symantec Enterprise Security (2011), www.symantec.com
Retrieved December 28, 2011

[18] **C. Wu,** "Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks" ,Expert Syst., 2009