International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 4, Issue 6 (Nov-Dec 2016), PP. 7-11

SPAM E-MAILS FILTERING TECHNIQUES

[1]Dipika Somvanshi, [2]Kanchan Doke
[1]Student, [2]Professor
[1,2] Computer Department, Bharati Vidyapeeth College of Engineering, Kharghar, Navi Mumbai
[1] dpksomvanshi@gmail.com

***Abstract*— Electronic mail is one of today's most important ways to communicate and transfer information. Because of fast delivery and easy to access, it is used almost in every aspect of communication in work and life. The continuous growth of email users has resulted in the increasing of unsolicited emails also known as Spam. SPAM email is well known problem for both corporate and personal users of email. Although SPAM has been well studied, both formally and informally, SPAM continues to be a significant problem. In current scenario, server side and client side anti spam filters are introduced for detecting different features of spam emails. So, Separation of spam from normal mails is essential. This paper surveys different spam filtering techniques. Techniques to separate spam mails are word based, content based, machine learning based and hybrid. Among them Machine learning techniques are most popular because of high accuracy and mathematical support. In this paper the overview of existing e-mail spam filtering methods is given. The classification, evaluation, and comparison of various machine learning-based methods are provided.**

***Keywords*— Spam Emails, Filtering, Traditional Methods, Machine Learning Based Methods, Content based filters.**

I. INTRODUCTION

Spam is the use of electronic messaging systems (including most broadcast media, digital delivery systems) to send unsolicited bulk messages indiscriminately. In this article it is considered the e-mail spam. E-mail spam, also known as junk e-mail or unsolicited bulk e-mail (UBE), is a subset of spam that involves nearly identical messages sent to numerous recipients by e-mail. Day by day the amount of incoming spam increase and, scammer attacks are becoming targeted and consequently more of a threat. Spam, which is unsolicited bulk email, has packed into everyone's daily life for decades. Researches even brought out a dreadful fact that all kinds of spam emails can account for as high as 88%~92% of all the emails delivered daily [1]. The topics of spam email vary from illegal products and services to intimidation and fraud, besides spam emails usually bring about potential risks like information theft by helping plenty of malware spread with extreme rapidity. Hence, in order to prevent the situation from deteriorating many solutions proposed and spam filtering technologies have been developed deeply and commercialized for years. But there still are hundreds of spam emails being encountered by every email user per year, indicating that the improvement is still necessary and urgent in spam filtering. Because spam emails could be exposed in every part of the email delivering process, there are many methods frequently used in spam filtering and always worked in conjunction, such as white lists/blacklists, challenge-response, rule-based filtering, keyword-based filtering, content-based filtering, etc [1].

This paper surveys various spam filtering techniques and then study and compares their performance, efficiency and speed of these techniques. Section II describes various spam filtering techniques. In section III, we are going to compare different machine learning techniques.

II. SPAM FILTERING TECHNIQUES

In this section various spam filtering techniques are discussed. They are mainly classified in four classes as, list based filters, content based, machine learning based, hybrid methods.

A. MACHINE LEARNING TECHNIQUES

Machine learning techniques aim to avoid the human efforts required to maintain rule based filters by automatically deriving a HAM/SPAM classifier. By definition, these techniques need to be fed already classified training data. Strong mathematical background is the reason behind the success of these techniques.

1. Clustering techniques:

Clustering is a class of technique used to segregate objects or case observations into relatively comparable groups

called clusters. It classifies object or observations in such a manner so that objects in a same group are more similar to each other than to those in other group. Two examples of clustering techniques which have been applied to SPAM classification are K-nearest neighbors (KNN) and density-based clustering.

K- Nearest Neighbor: The working of this technique resolves around concept called 'characteristics vector'. The characteristic vectors are measure of similarities among all messages. Any new mails are classified in any of spam or ham class on basis of distance of that mail from both classes K-nearest neighbors (KNN) clustering indexes and converts emails to a high-dimensional vector and then measures the distance between the vectors of each email. Clusters are formed of neighboring, i.e., relatively close vectors. Once clusters have been formed SPAM classification need only be performed for a subset of any cluster population, as the result can then be inferred to apply to the other members of the cluster [1].

Density based clustering is another form of document clustering that has also been applied to SPAM classification. A

claimed advantage is the ability to process hashed versions of messages, thus preserving user privacy. These methods depend on having sensitive comparators. These comparators are usually either fast or sensitive. The challenge is finding comparators that are both sufficiently fast and sufficiently sensitive [6].

2. Naïve Bays Classifier Method:

Baysian filters, considered the most advanced form of content based filtering, employs laws of mathematical probability to determine which messages are legitimate and which are spam. In order to block spam, end user must "train" emails manually by flagging each message legitimate & spam.

The percentage of false positive generated by Bayesian filters are low, and they are self-adapting to stop new SPAM by receiving ongoing training form the user. Spamicity of a word is the probability of a word being spam. For calculating spamicity of total message, this technique lets us combine the probability of multiple independent events into one number. Each word has particular probabilities of occurring in spam email and in legitimate email. The overall process has the following steps:

Train the filter.

Calculate the probability of words.

Combine the word probability to classify mail.

At the start, emails are manually classified as spam or ham. After training, the probability of each word in the spam and legitimate mail is calculated by the following formula. Then this data is stored by the spam filter in its database. Filter also maintains a spam token database and the nonspam token database which contains count of each word in email.

$$Pr(S/W) = \frac{Pr(W/S)}{Pr(W/S)+Pr(W/H)}$$

Where,

Pr(S/W) = probability that a message is a spam.

Pr (W/S) = probability that the specified word appears in a spam message.

Pr (W/H) = probability that specified word appears in ham message.

The overall probability of email is calculated as

$$Pr(S/W) = \frac{Pr(W/S).\ Pr(S)}{Pr(W/S).\ Pr(S) + Pr(W/H).Pr(H)}$$

Where,

Pr (H) = probability that any given message is ham.

Disadvantages of this technique are

Words which occur in spam are misspelled.

Spammers insert sensitive words in the form of images in a spam mail and Bayesian Classifier can't analyze images [1].

3. Support Vector Machine Method:

SVMs are a supervised learning method, used in text classification, that have more recently been applied to the SPAM identification problem. The Support Vector Machine is one of the most modern techniques used in mail classification. In abstract view, it is a kernel machine with the strong mathematical base. It is a technique of pattern recognition and data analysis. The training sample is a set of vectors of n attributes. At the end of training phase, we can say that, we are in hyperspace having dimensions equal to the number attributes. In process of spam filtering, SVM builds hyperspace with two classes, namely, spam and ham. These two classes are separated by a hyperplane. Every mail instance is treated as a single point with n dimensions in hyperspace. The distance between the hyperplanes and points of each class, is kept maximum, for good separation [1]. Here in fig. 2, Plane1 is good classifier and Plane2 doesn't classify all instances. It may also happen that, we can't find good separator hyperplane (Plane 1) as in fig. 2. In such case, hyperspace is called as non-linearly separable. To obtain linear separation in the non-linearly separable hyperspace, it is extended to more dimensions. SVM method considers only the nearest points in hyperspace to find hyperplane [1].

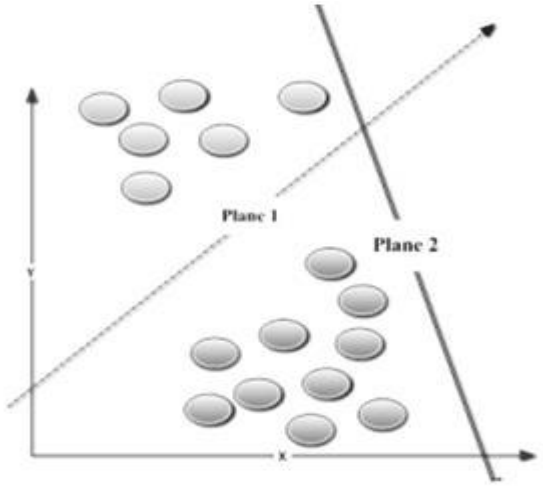

Fig.2. Hyperplane that separate the two classes

4. Neural Network Method:

Multilayer Perceptron (MLP) has been used in this paper. Neural network is based on biological nervous systems such as brain. It is useful for problem for which no algorithmic solution is available or algorithmic solution is too complicated to be exemplified. Neural networks are good at solving problem that can be solved by human more efficiently rather than compute. It can learn both linear and non linear relationship from data set. MLP uses back propagation in order to build a robust neural network. For each input to the MLP the resulting output is compared to desired output and error is calculated, this error is fed back to neural network and weights are adjusted such that it gives nearby results to the desired output on each iteration. Disadvantage of this technique is that due to difference in architecture from microprocessor's architecture, it needs to be emulated. Also it takes more processing time to build large neural network [1].

5. ID3

ID3 is a non-incremental algorithm used to build a decision tree from a fixed set of observations (dataset). The resulting tree is used to classify test observation. Each observation is represented by its features or attributes and a class to which it belongs. Representation of decision tree is as follows Leaf node of decision tree contains class name. Non leaf nodes are decision nodes. These nodes contain condition involving an attribute with sub branches having possible value for that attribute ID3 uses information gain measure to select decision node. Information gain indicates the ability of a given attribute to separate training examples into classes. Higher the information gain, higher is the ability of the attribute to separate training observation. Information gain uses entropy as a measure to calculate the amount of uncertainty in dataset. It builds fastest and short tree and considers attributes that are enough to classify data. But it suffers from over-fitting problem if training data is small [3].

6. J48

J48 builds decision trees from a set of training data using the concept of information entropy. J48 examines the normalized information gain that results from choosing an attribute for splitting the data. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 classifier recursively classifies until each leaf is pure, meaning that the data has been categorized as close to perfectly as possible. J48 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S = s1,s2,... of already classified samples. Each sample si = x1,x2,... is a vector where x1,x2,... represent attributes or features of the sample. The training data is augmented with a vector C = c1,c2,... where c1,c2,... represent the class to which each sample belongs. At each node of the tree, J48 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The J48 algorithm then recurs on the smaller sublists. This algorithm has a few base cases

. • All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.

• None of the features provide any information gain. In this case, J48 creates a decision node higher up the tree using the expected value of the class.

• Instance of previously-unseen class encountered. Again, J48 creates a decision node higher up the tree using the expected value [3].

7. Boosting

Boosting is a learning algorithm which is based on the idea of combination of many weak hypotheses, for example as in the AdaBoost system. A learner is trained in each stage of the classification procedure, and the output of each stage uses to reweight the data for the future stages [6].

Boosting algorithms with confidence rated predictions have been proposed as being well suited to the SPAM filtering problem, and that they can outperform both Bayesian and decision tree methods [6].

Maximum entropy models are another machine learning technique from natural language processing that has also been applied to SPAM filtering. Memory based learning are non-parametric inductive learning paradigm that stores training instances in a memory structure on which predictions of new instances are based and have also been applied to the problem of SPAM [1,6] .

B. LIST BASED TECHNIQUES

List based filters maintain the list of legitimate mail senders and spammers and allow access or block mail accordingly.

1. Blacklist:

This technique maintains the list of spammers. This kind of list is explored and maintained by the organization itself to the save a mail infrastructure. Blacklist contains user names or IP addresses of mail senders which normally send spam. Drawback of Blacklist approach is false positivity, in case, spammer sends spam from IP of a legitimate user [1].

2. Real Time Blackhole List:

The way of working of above and this approach is nearly same. Only difference is, the list of spammers is maintained by the third party here. This approach gives better results compared to above, because the list is updated frequently. To avail the facility of this type, organization has to subscribe to a third party. This can save manpower resources of organization to maintain the list. Lacunas of this approach, organization have less control over the list and false positivity [1].

3. Whitelist:

As a name suggests, this technique is exactly opposite of Blacklist. Unlike Blacklist, it makes the list of legitimate senders only. It is suitable when numbers of mail senders are fixed. The serious drawback of this is new sender, who is not a spammer, can't send a mail. To achieve this list needs to be updated. Normally, the mails are separated in good, bad and unknown categories. The popular tool which works on this technique is Spamhaus [1].

4. Greylist:

As compared to the above three techniques this is newer. Work of it is based on an idea that, spammers send a batch of spam mails only once. So the spams can be avoided by not receiving mail in inbox at first attemp. Here, receiving mail server having Greylist not accepts any mail in the first attempt. It sends failure message to generating mail server. If the sender

is not spammer, then sending mail server sends mail again, at this time it is accepted by receiving mail server and ultimately by the receiver. If a sender is spammer, it will not resend the mail. This technique is very effective to avoid spam. Benefit is infrastructure requirement is very marginal. Greylist approach causes delay in delivery of mail, so not suitable in case of urgency [1].

C. CONTENT BASED TECHNIQUES

It is a very popular technique to avoid spam, in which mails are evaluated for words or phrases to determine mail as a spam or legitimate.

1. Word Based:

It blocks mail as a spam, if mail has certain word having spamicity character. Mostly, spams contains the terms which are rarely used in legitimate mails. So, it is easy to block the spams. One serious problem is that if a filter is configured to block mails containing more common words then it increases the false positivity. The list of such words is available online [1].

2. Heuristic filters:

It outperforms the normal word based filter. The word heuristic refers to some intuitive criteria, rather than simple technical metrics. Mostly point and score is criteria to classify the mail as legitimate or spam. More points are assigned to words or a term which occurs frequently in spam. The terms frequently used in legitimate mails are assigned with low score. At the end, a score of mail is calculated. If the score is beyond some predefined threshold, it is marked as spam. Experienced spammers can easily bypass this type of filter by avoiding use of terms with the high score. Also Heuristic filters make use of various algorithms to examine the email [1].

D. OTHER OR HYBRID TECHNIQUES

1. Challenge/Response system:

This method normally works in three steps as shown in fig. 1

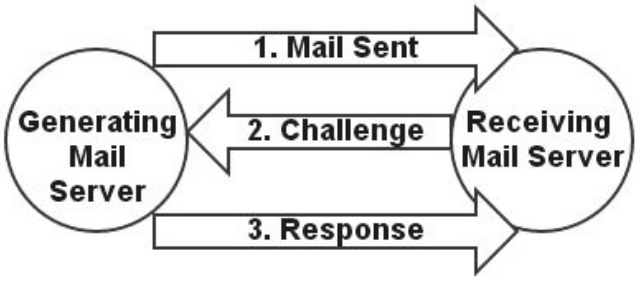

Fig.1. Sample challenge/ response environment

In the first step generating mail server sends mail to receiving mail server. Then in the next step receiving mail server sends then mails from it are marked as spam and not delivered to challenge to generating mail server. Finally, in the third step, generating mail server sends response to the challenge. Challenge is normally a random number generated by the receiving mail server. Spammers generally fail to fulfill the challenge because; spammers have to response all challenges, which is not a possible or tedious job. If a sender solves it, current mail and all future mails from that sender are marked as legitimate [1].

2. Collaborative filter:

It is community based approach. Users form the 'community' to decide the nature of the mail. Users in community mark mail as spam or legitimate by setting some flag. Then its record is kept in the central database. If flag count of some mail goes beyond some predefined threshold limit, mail is marked as spam and blocked from reaching other users. Joseph S. Kong highlighted some benefits of collaborative filter. SpamNet filter good example of this category. One drawback of it is, every time it builds a new community by ignoring the existing ones. If an existing community contains spammers, then they don't increment flag count, results in spam can be marked as legitimate mail [1].

3. Image Based Filtering:

All above mentioned techniques fail to identify spam, which have an image in it; they just decide spam by words or phrases in mail. So savvy spammers can easily send spam mails, by adding spam message in the form of image to mail. To detect such spams, an image based spam filters with Optical Character Recognition (OCR) is used by Wanli Ma, et.al. OCR extracts text from an image, and then normal approach is followed. Point to concern here is, misrecognition of OCR is unavoidable, and accuracy of this filter is directly depends on accuracy of OCR. To improve overall performance it, Markov Model which tolerates misspells is embedded with the filter [1].

4. SMTP Server Based Filter:

Spam mail can be easily detected at SMTP server. Tools like ASSP and Xeams, filters spam at SMTP server itself. Xeams supports SMTP, IMAP, POP3 services, which facilitate easy spam filtering. Xeams has spam filter having accuracy near about 99%. ASSP is a hybrid approach which uses Blackhole Lists and Bayesian filtering approaches [1].

5. DNS Blackhole List (DNSBL) check:

In this technique receiving mail server keeps the DNS based list of blacklisted mail servers. The IP address of sending mail server is checked against DNS based blacklist. In case, if a sender is blacklisted then mails from sender are marked as a spam and not delivered to the receiver. To know IP of our mail server MXtoolbox is used. The list of mail servers in DNSBL is available online [1].

6. DNS Lookup Systems:

DNS protocol is used over the internet for domain name to IP address mapping. Mail servers use DNS protocol to identify themselves. This protocol can be used effectively to spam filtering. The Mail Exchange (MX) record is available with

every valid mail. Based on MX record, DNS server verifies the name of mail server (domain name) as valid or not. If the domain name doesn't have the valid MX record then the mails sent from it are treated as spam and not delivered to receivers [1].

7. Bag of Words model:

Bag of Words approach is used widely in Natural Language Processing (NLP) and Information Retrieval (IR). It is broadly used for the document classification. Herein this approach, each word in mails is listed in the document and then associated with an index. Then every mail is represented as a vector of size equal to the number of words in that mail, and value in each vector is, count of the corresponding word in mail. Here document or dictionary formed is termed as 'Bag'. This model can be compared with Naïve Bayes classifier, where two Bags, one for legitimate mails and other for spam mails is maintained.

III. COMPARISON OF MACHINE LEARNING TEHNIQUES

TECHNIQUE	ACCURACY	TIME TAKEN	TP	TN	FP	FN
SVM	93.617	l.92 sec	0.88	1	0	0.12
Naïve Bays	9l.4894	0.46 sec	0.84	1	0	0.16
Clustering	53. 19 15	1.33 sec	1	0	1	0
ANN	-	More time required	-	-	-	-
Decision Tree	93.6 17	0.81 sec	0.88	1	0	0.12

Table 1 Performance Comparison Of Classification Techniques

From the above comparison table it can be concluded that from all techniques that have been used here, Naïve Bays technique gives faster result and good accuracy over other techniques (except SVM and Decision Tree). SVM and Decision Tree give better accuracy than Naïve Bays but take more time to build a model. There is a trade-off between time and accuracy. So which technique is used depends on the application at hand [2].

Conclusion

Summarizing above-listed, we obtain the following conclusions. We have presented a wide range of the techniques that have been used or proposed for use to fight SPAM, and attempted to indicate which SPAM filters use which techniques. We have sought to arrange these techniques in an orderly and informative manner, in the hope that the result will prove helpful in the continuing fight against SPAM, by allowing intelligent selection of SPAM filters by practitioners, and more consistent and informed treatment of SPAM filters in the academic literature compared with the previous situation. The taxonomy presented here is clearly preliminary in nature, and non-exhaustive. A clear task for the future is to expand it with information about additional SPAM filters and techniques, and to address any refinements that become apparent during that process.

REFERENCES

[1] "Spam Filtering Techniques & Map Reduce with SVM", Amol G. Kakade, Prashant K. Kharat, Anil Kumar Gupta, Tarun Batra, Department of Information Technology,2014 Asia-Pacific Conference on Computer Aided System Engineering (APCASE), 978-1-4799-4568-9/14/$31.00 ©2014 IEEE.

[2] "A Survey and Evaluation of Supervised Machine Learning Techniques for Spam E-Mail Filtering", Tarjani Vyas, Payal Prajapati, & Somil Gadhwal, Institute Of Technology, IEEE international conference on Electronics, Computers & Communication Technologies, 978-1-4799-608S-9/1S/$31.00©201S IEEE

[3] " MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION ", W.A. Awad1 and S.M. ELseuofi, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011

[4] " Spam Mails Filtering Using Different Classifiers with Feature Selection and Reduction Techniques", Amit Kumar Sharma, Renuka Yadav, ICFAI University, 2015 Fifth International Conference on Communication Systems and Network Technologies, 978-1-4799-1797-6/15 $31.00 © 2015 IEEE

[5] " Survey on Spam Filtering Techniques", Saadat Nazirova , Institute of Information Technology of Azerbaijan National Academy of Sciences, Communications and Network, 2011, 3, 153-160

[6] "A TAXONOMY OF EMAIL SPAM FILTERS", HASAN SHOJAA ALKAHTANI, PAUL GARDNER-STEPHEN, AND ROBERT GOODWIN, Computer Science Department, College of Computer Science and Information Technology, King Faisal University, Saudi Arabia

[7] "Origin (Dynamic Blacklisting) Based Spammer Detection and Spam Mail Filtering Approach", Nikhil Aggrawal, Shailendra Singh, Dept. of Computer Science, IEEE international paper on computer & Networks, ISBN: 978-1-4673-9379-9 ©2016 IEEE

[8] "EVALUATION OF DECEPTIVE MAILS USING FILTERING & WEKA", Sujeet More, Ravi Kalkundri, IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS'15, 978-1-4799-6818-3/15/$31.00 © 2015 IEEE

[9] "Spam Filtering by Semantics-based Text Classification ",Wei Hu, Jinglong Du, and Yongkang Xing, 8th International Conference on Advanced Computational Intelligence, Chiang Mai, Thailand; February 14-16, 2016, 978-1-4673-7782-9/16/$31.00 ©2016 IEEE

[10] "Spam Classification Based on Supervised Learning by Machine Learning Techniques", Ms. D. Karthika Renuka, Dr. T. Hansapriya,Mr. M. Raja Chakravarthi, Ms. P. Lakshmi Surya, IEEE international paper on Data mining, 978-1-61284-764-1/11/$26.00 ©2011 IEEE