Spam Mail Filtering: A Survey of Effective Spam Mail Detection Techniques

Renuka Yadav[1] Jignesh Vania[2]

[1]P.G. Student [2]Assistant Professor

[1,2] Department of Computer Engineering

[1,2] L.J.I.E.T Ahmedabad, India

Abstract— with the continuous growth of email users has resulted in the increase of unsolicited emails also known as Spam. E-mail spam, known as unsolicited bulk Email (UBE), junk mail, or unsolicited commercial email (UCE), is the practice of sending unwanted e-mail messages, frequently with commercial content, which is normally in large quantities to set of recipients. There are many spam filters available that uses various approaches to identify the incoming message as spam or ham, ranging from white list / black list, Bayesian analysis, keyword matching, mail header analysis, postage, legislation, and content scanning etc. but we are still flooded with spam emails every day. This is not because the filters are not powerful enough, it is due to the swift adoption of new techniques by the spammers and the inflexibility of spam filters to adapt the changes. In our work, we employed supervised machine learning techniques to filter the email spam messages. Widely used supervised machine learning techniques namely C 4.5 Decision tree classifier, Multilayer Perceptron, Naive Bayes Classifier are used for learning the features of spam emails and the model is built by training with known spam emails and legitimate emails. The results of the models are discussed.

In current, server side and client side anti-spam filters are introduced for detecting different features of spam emails. However, recently spammers introduced some new tricks consisting of embedding spam contents into digital image, pdf and doc as attachment which can make ineffective to current techniques that is based on analysis digital text in the body and subject fields of email.

Key words: Spam, Naïve bayesian filter, image spam, spam detection, machine learning, stemming.

I. INTRODUCTION

In this modern era where information exchange through email service is popular and common, it comes with an opportunity for the frauds that are easily possible through this technology. Spam Email is that such technique usually deployed for fraudulent. Spam mail stands for "Self-Promotional Advertising Message" which can be used for advertising, gaining personal information of end users or getting account credentials etc which will affect the end user. At the same time it is now-a-days getting popular as "unsolicited bulk mail". Email spam causes system to experience overload in its bandwidth and also degrades the server storage capacity. Also, phishing spam mails are becoming a serious security threat to the end users as they provide their personal information and account credentials. It is responsible for the wastage of user time, economic loss; extend virus, Trojans, loss of work productivity [2]. Hence, there is a serious need to implement spam mail filters which can be deployed either at client side or server side.

Spam filtering techniques are classified to segregates ham and spam mails. Filtering techniques majorly focus on three levels which are subject, email address and contents of the message that is being sent [2]. Email addresses and subject of messages have technologies such as: pattern matching, black and white listing. But spammers can easily bypass such filters and can result in increased rate of false negatives which could be depressing.

Content based spam filtering is one of the most effective solutions foe spam detection. It is based on the feature selection and text classification methods such as Naïve Bayesian classifier, Random forest and SVM etc.

However, another trick is also applied by spammer that is image spam. Its concept is to embed text onto the image and send it in attachment so that it will bypass the filter. OCR provides the facility to extract the text embedded onto image. But spammers fool this technique by using text in the form of CAPTCHA and its hard challenge for OCR [3].

So the problem is that bulk mails are being received in our mailbox without our knowledge that degrades system performance. So there is need to secure the system by detection and prevention from such spam mails.

II. BACKGROUND

Problem of getting undesired mails is becoming a serious issue day by day. Spammers are usually more active at the time festivals and sales as the end users generally look for the festive offers and sales and hence get trapped this way.

A. Sources of Spam

1) Directory Harvest Attack [4]

In this technique, spammers collect the email addresses of end users from articles, mailing list, forums etc.

2) Botnet [4]

This technique includes inserting malicious code or malware that can turn the system into a zombie computer.

3) Internet Hoaxes and Chain Letters [4]

Since our society have people with different behavior and psychology so spammers attack on people psychology and send them fake stories such as "if you send this to 10 people, a miracle will happen to you within 2 days".

4) Social Networking

Social networking is way to connect people globally. So, it's a best way to get the personal information from there. Phishing attacks are the best way of this kind.

5) Backscatter
This technique allows the spammers to send bounce mails in which the "FROM" and "TO" fields contain a valid email address.

B. Types of Spam

1) Blank Spam
A spam that contains no subject line, body or advertisement. Because its nature is unsolicited and mailed in bulk so it suits for spam mail.

2) Attachment Spam
Spammers are using Zip file, PDF attachments and excel to bypass the traditional spam filters.

3) Email Scam
Such mail consists of fake stories like "your profile has been shortlisted for a well reputed MNC in Delhi NCR".

4) Image Spam
Putting the spam message into image attached with message.

Fig. 1: Example of image spam mail [7]

5) SMS Spam
Mobile spam is becoming frequent means of spreading advertisement, date services, premium services, coaching classes, health centers etc. But most of the spam filters can effectively solve the problem of mobile spam.

6) Trojan horse Email
Such email offers a photograph or patch for software vulnerability.

7) Phishing mails
Such mails look very similar to mails from bank and spammers tries to get users account credentials.

C. SPAM FILTERING

1) Origin Based Spam Filtering [5]
Such filtering is applied before receiving the message. It collects the spam senders address list and blacklists it if sender email address is in the black list.

2) Filtering Based on Traffic Analysis [5]
In this, log files of SMTP server can be used to find the duplicate mail message and anomalies in the normal traffic.

3) Rule Based Spam Filtering [5]
Some set rules have to be established which is quite lengthy process and that will determine whether incoming message is spam or ham.

4) Challenge Response Filter [5]
This technique is not very much effective because of its working technique. Here, the sender have to authenticate itself as after sending the mail it will get a response mail to validate itself. It fails in case when the emails are machine originated mails as the machines don't have the capability to revert such mail.

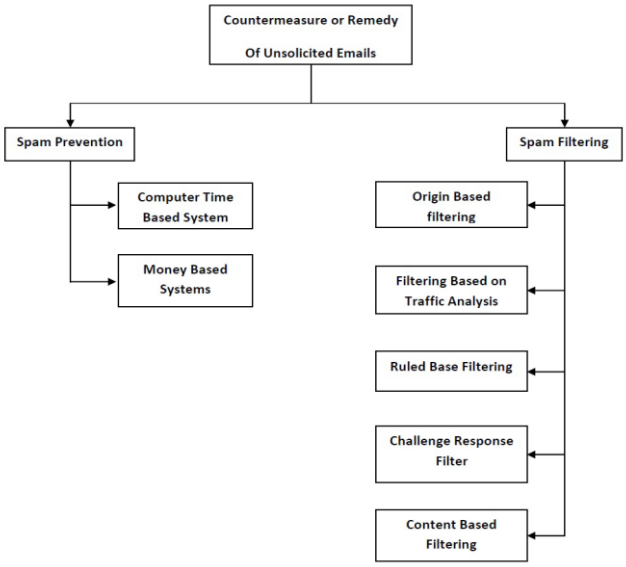

Fig. 2: Remedy for Spam Filtering [6]

D. Methodology

Most of the spam filtering techniques is based on text categorization methods. Thus filtering spam turns on a classification problem. In our work, rules are framed to extract feature vector from email. As the characteristics of discrimination are not well defined, it is more convenient to apply machine learning techniques. Three machine learning algorithms, C4.5 Decision tree classifier, Multilayer perceptron and Naive bays classifier are used for learning the classification model.

1) Multilayer Perceptron
Multilayer Perceptron (MLP) network is the most widely used neural network classifier. MLP networks are general purpose, flexible, nonlinear models consisting of a number of units organized into multiple layers [1]. The complexity of the MLP network can be changed by varying the number of layers and the number of units in each layer. Given enough hidden units and enough data, it has been shown that MLPs can approximate virtually any function to any desired accuracy. MLP is becoming a popular tool in case when there is almost no or less knowledge about the relation between input values and their respective outputs.

2) Decision Tree Classifier
Decision Tree Classification generates the output as a binary tree like structure called a decision tree, in which each branch node represents a choice between a number of possible alternatives, and each leaf node will represent a classification or decision [1]. A Decision Tree model contains a set of rules to predict the value of target variable. This algorithm basically performs well, even where there are varying numbers of training inputs and considerable numbers of attributes in large databases. J48 algorithm is an implementation of the C4.5 decision tree learner. This implementation produces decision tree models. To implement the decision tree for the purpose of classification this algorithm uses greedy approach. A decision-tree model is built by analyzing training data and the model is used to classify test data. J48 generates decision trees, the nodes of

which evaluate the existence or significance of individual features.

3) Naïve Bayesian Classifier

The naive bayes classification technique (NB) is quite simple but effective classifier. It has been used in a lot of applications for information processing which includes natural language processing, information retrieval, etc [1]. The Naive Bayes' Classifier technique is completely based on Bayesian theorem and is best suited for a scenario in which the dimensionality of the inputs is quite high. Naive Bayes' classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. The Naive-Bayes' algorithm computes conditional probabilities of the classes given the instance and picks the class with the highest posterior. Depending on the precise nature of the probability model, naive Bayes' classifiers can be trained very efficiently in a supervised learning setting.

4) Concept Drift

Evolving changes in pattern of email content both for spam and ham are able to bypass the existing filters. These changes make spam filters obsolete. So, detecting these changes and updating the model regularly is an important issue.

The learned model of a classifier should adapt itself to classify new emails correctly. This problem is concept drift and can be applied to find critical solutions in which the model should be updated.

To determine concept drift, two factors are recently considered; firstly label of data and secondly user feedback. It can be estimated that concept drift occurs when the information distribution between two consecutive blocks is very different [8].

5) Probabilistic boosting tree for image spam

Image spam is new techniques to bypass existing filters. In such spam, text is embedded onto images and sent either in body of email or as an attachment. One approach to deal with such spam mails is probabilistic boosting tree which determines whether an incoming mail is spam or not based on the global image features such as color and gradient orientations histogram. Another approach is to use OCR tool to extract the text embedded onto the images and then using traditional spam filters [9].

6) Behavior and keyword stemming

Spammers are usually active during the non-office hours because at that time, there is high availability of larger bandwidth and hence chances of remaining undiscovered are high.

Two main factors or behavioral features identified are the presence of the URL (Uniform Resource Locator) and also, the time at which the mail was dispatched.

Stemming is function proposed by Porter to find the root or core word from the given word so that the most common keywords can be found in spam and ham mails [10].

Based on three criteria, the performance of three models has been evaluated. These criteria are prediction accuracy, correctly classified instances and false positive rate [1]. Performance has been evaluated using 10-cross validation technique. Total 1500 instances are taken out of which 750 instances are spam and rest is legitimate mails. The results show that Multilayer Perceptron produces better result than other algorithms. Table. 1 shows the performance analysis of naïve Bayesian, J48 and MLP classifiers.

Evaluation / Criteria	Naive Bayes	J4.8	MLP
Training time (in sec)	0.15	0.20	138.05
Correctly Classified Instances	1479	1449	1490
Prediction Accuracy (%)	98.6	96.6	99.3
False Positive (%)	5	4	1

Table. 1: Comparison of Various Filtering Algorithms

III. PROBLEM STATEMENT

A big problem is bulk of spam emails that are received in our mailbox without our knowledge and target to affect the systems. Researchers have proposed many suggestions, many firms and developers have developed several tools and applications to prevent these emails. But still the problem is occur whenever spam filters are introduced, the spammer try to find out the way of bypass entry in the system and generate more tricks to affect the system. Today's mostly spammer sends spam mails in image format and uses several tricks in images.

There is find out some problems during the literature survey. These are:

1) How to make more effective to spam filters.
2) Better identify spam emails and classification of ham and spam emails from large data.
3) How to better identification of image spam emails from large volume of data.
4) Understand human vision that used by spammer.
5) Find out new tricks which applied by spammer to send image spam emails.

IV. CONCLUSION

Researchers are trying to implement new spam filters that can prevent to destination by blocking at client level or server level. Naïve Bayesian algorithm works well when there is high dimensionality in input data whereas multilayer perceptron performs better when the knowledge about input variable is quite less but it takes much more time to produce results compared to others. Furthermore, more research work can be done by combining the methods such as feature extraction, feature selection, word frequency count, classification etc to produce comparable results.

REFERENCES

[1] Christina, V., S. Karpagavalli, and G. Suganya, "Email spam filtering using supervised machine

learning techniques", *Journal of Computer Science and Engineering* 2.9 (2010): 3126-3129.
[2] Caruana, Godwin, and Maozhen Li, "A survey of emerging approaches to spam filtering", *ACM Computing Surveys (CSUR)* 44.2 (2012): 9.
[3] Fumera, Giorgio, Ignazio Pillai, and Fabio Roli. "Spam filtering based on the analysis of text information embedded into images." *The Journal of Machine Learning Research* 7 (2006): 2699-2720.
[4] Fong, Michael. "Spam or Ham." (2008).
[5] Wu, Jiansheng, and Tao Deng. "Research in Anti-Spam Method Based on Bayesian Filtering." *Computational Intelligence and Industrial Application, 2008. PACIIA'08. Pacific-Asia Workshop on*. Vol. 2. IEEE, 2008.
[6] Garcia, Flavio D., Jaap-Henk Hoepman, and Jeroen Van Nieuwenhuizen. "Spam filters analysis." *Security and Protection in Information Processing Systems*. Springer US, 2004. 395-410.
[7] Gao, Yan, Alok Choudhary, and Gang Hua. "A comprehensive approach to image spam detection: from server to client solution." *Information Forensics and Security, IEEE Transactions on* 5.4 (2010): 826-836.
[8] Xiaofeng, Qiu, Hao Jihong, and Chen Ming. "Flow-based anti-spam." *IP Operations and Management, 2004. Proceedings IEEE Workshop on*. IEEE, 2004.
[9] Attar, Abdolrahman, Reza Moradi Rad, and Reza Ebrahimi Atani. "A survey of image spamming and filtering techniques." *Artificial Intelligence Review* 40.1 (2013): 71-105.
[10] Wikipedia. Types of email spams. URL http://en.wikipedia.org/wiki/Email_spam.

