Email Spam Filtering using Supervised Machine Learning Techniques

V.Christina[#], S.Karpagavalli[*], G.Suganya[#]

[#]M.Phil Research scholar Department of Computer Science(PG)
P.S.G.R Krishnammal College for Women

[*]Senior Lecturer
GR Govindarajulu School of Appiled Computer Technology

***Abstract*— 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, in large quantities to an indiscriminate set of recipients. Spam is prevalent on the Internet because the transaction cost of electronic communications is radically less than any alternate form of communication. There are many spam filters using different approaches to identify the incoming message as spam, ranging from white list / black list, Bayesian analysis, keyword matching, mail header analysis, postage, legislation, and content scanning etc. Even though we are still flooded with spam emails everyday. 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, Naïve 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.**

***Keywords*— Spam, Spam filter, Spammer, Mail header, Machine learning, Classifier**

I. INTRODUCTION

The internet has become an integral part of everyday life and e-mail has become a powerful tool for information exchange. Along with the growth of the Internet and e-mail, there has been a dramatic growth in spam in recent years. Spam can originate from any location across the globe where Internet access is available. Despite the development of anti-spam services and technologies, the number of spam messages continues to increase rapidly. In order to address the growing problem, each organization must analyze the tools available to determine how best to counter spam in its environment. Tools, such as the corporate e-mail system, e-mail filtering gateways, contracted anti-spam services, and end-user training, provide an important arsenal for any organization. However, users cannot avoid the very serious problem of attempting to deal with large amounts of spam on a regular basis. If there are no anti spam activities, spam will inundate network systems, kill employee productivity, steal bandwidth, and still be there tomorrow.

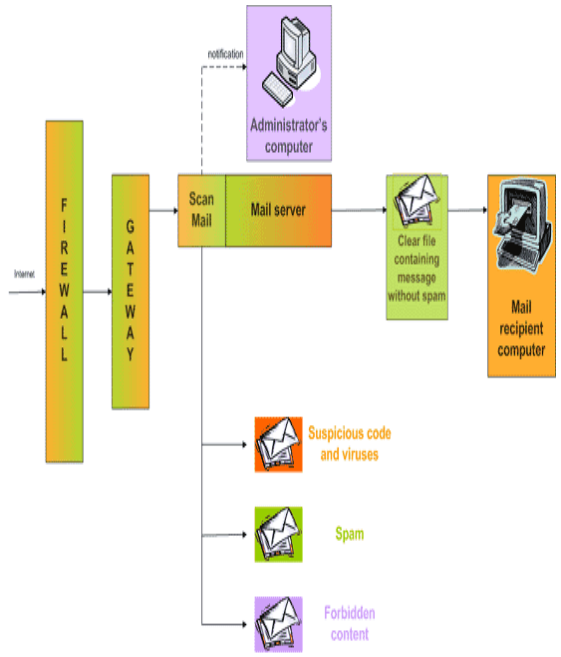
II. SPAM FILTER ARCHITECTURE AND METHODS

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, in large quantities to an indiscriminate set of recipients. The technical definition of spam is 'An electronic message is "spam" if (A) the recipient's personal identity and context are irrelevant because the message is equally applicable to many other potential recipients; and (B) the recipient has not verifiably granted deliberate, explicit, and still-revocable permission for it to be sent'. The risks in filtering spam are sometimes legitimate mails may be rejected or denied and legitimate mails may be marked as spam. The risks of not filtering spam are the constant flood of spam clogs networks and adversely impacts user inboxes, but also drain valuable resources such as bandwidth and storage capacity, productivity loss and interfere with the expedient delivery of legitimate emails.

Spam filters can be implemented at all layers, firewalls exist in front of email server or at MTA(Mail Transfer Agent), Email Server to provide an integrated Anti-Spam and Anti-Virus solution offering complete email protection at the network perimeter level, before unwanted or potentially dangerous email reaches the network. At MDA (Mail Delivery Agent) level also spam filters can be installed as a service to all of their customers. At Email client user can have personalized spam filters that then automatically filter mail according to the chosen criteria. Figure 1. shows the typical architecture of spam filter.

The several different methods to identify incoming messages as spam are, Whitelist/Blacklist, Bayesian analysis, Mail header analysis, Keyword checking. A whitelist is a list, which includes all addresses from which the users always wish to receive mail.

User can add email addresses or entire domains, or functional domains. An interesting option is an automatic whitelist management tool that eliminates the need for administrators to manually input approved addresses on the whitelist and ensures that mail from particular senders or domains are never flagged as spam.



The number of records can be configured. When an overflow occurs, obsolete records are overwritten. A blacklist works similarly to competitive alternatives: this is a list of addresses from which user never want to receive mail. Mail header checking consists of a set of rules that, if a mail header matches, triggers the mail server to return messages that have blank "From" field, that lists a lot of addresses in the "To" from the same source, that have too many digits in email addresses (a fairly popular method of generating false addresses). It also enables to return messages by matching the language code declared in the header.

In Bayesian analysis, the word probabilities (also known as likelihood functions) are used to compute the probability that an email with a particular set of words in it belongs to either category. This contribution is called the posterior probability and is computed using Bayes' theorem. Then, the email's spam probability is computed over all words in the email, and if the total exceeds a certain threshold, the filter will mark the email as a spam. Keyword checking is another method widely used in filtering spam. It works by scanning both email subject and body. Using "conditions" i.e. combinations of keywords is a good solution to enhance filtering efficiency. We can specify combinations of words and update the list that must appear in the spam email. All messages that include these words will be blocked.

III. 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, C 4.5 Decision tree classifier, Multilayer perceptron and Naïve bayes classifier are used for learning the classification model.

A. 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 organised into multiple layers. 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. In other words, MLPs are universal approximators. MLPs are valuable tools in problems when one has little or no knowledge about the form of the relationship between input vectors and their corresponding outputs.

B. C 4.5 Decision Tree Induction

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 alternatives, and each leaf node represents a classification or decision. A Decision Tree model contains rules to predict the target variable. This algorithm scales well, even where there are varying numbers of training examples 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. The algorithm uses the greedy technique to induce decision trees for classification. A decision-tree model is built by analyzing training data and the model is used to classify unseen data. J48 generates decision trees, the nodes of which evaluate the existence or significance of individual features.

C. Naïve Bayes Classification

The naive bayes classifier (NB) is a simple but effective classifier which has been used in numerous applications of information processing including, natural language processing, information retrieval, etc. The Naive Bayes Classifier technique is based on Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Naïve 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 inducer 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.

IV. Feature Extraction

The work is based on rules and uses a score-based system. The rules are framed by analyzing the mail header information, keyword matching and the body of the message. And a relative score is assigned to each rule.

There are number of rules framed by considering the various features that will aid to identify the spam messages effectively. Each rule performs a test on the email, and each rule has a score. When an email is processed, it is tested against each rule. For each rule found to be true for an email, the score associated with the rule is added to the overall score for that email. Once all the rules have been used, the total score for the email is compared to a threshold value. If the score exceeds the threshold, then the email is marked as spam and the others are classified as legitimate mail. In this work, the rules used are

TABLE I
SCHEME OF RULES ASSIGNED TO EACH SPAM FEATURE

From name meaningful
From domain name
Blocked IP
Apostrophe in From name
From name in Auto Whitelist (AWL)
From address in User's Block list
From address in User's White list
Content Type
Content Boundary exists
To name meaningful
To address Undisclosed recipients
To header original
From address and To address same
Is subject present
Subject content has obfuscate words
Is forwarded message
Is reply message
Subject Reply without reference header
Is message body exists
Sensual message
Repeated double quotes in body
Character set includes foreign language
More blank lines in body

In these 23 rules, some are simple and some are associated with one another. A simple rule could search for a word 'Viagra' in subject line of an email, while a complex rule may involve comparing an email against an online database of spam. Each rule adds to the overall score, so an email that triggers only one rule due to the use of the word 'Viagra' will not necessarily mark an email as spam. However, if an email triggers several rules, it will have a combined score that could be over the threshold and the mail could be marked as spam.

V. EXPERIMENT AND RESULTS

The email spam filtering has been carried out using WEKA. The Weka, Open Source, Portable, GUI-based workbench is a collection of state-of-the-art machine learning algorithms and data pre processing tools.

The training dataset, spam and legitimate message corpus is generated from the mails that we received from our institute mail server for a period of six months. The mails are analyzed and 23 rules are identified that extremely ease the process of classifying the spam message. The corpus consists of 750 spam messages and 750 legitimate messages. From the corpus, the feature vectors are extracted by analyzing message header, keyword checking, whitelist/blacklist etc. The class labels are designated as L and S to represent legitimate and spam message respectively.

The machine learning techniques Naïve Bayes Classifier, C 4.5 Decision tree classifier, Multilayer Perceptron are used for training the dataset in WEKA environment.

The training is carried out with the feature vectors extracted by analyzing each message header and keyword checking and whitelist/blacklist.

The performance of the trained models is evaluated using 10-fold cross validation for its predictive accuracy. Predictive accuracy is used as a performance measure for email spam classification. The prediction accuracy is measured as the ratio of number of correctly classified instances in the test dataset and the total number of test cases. In spam filtering, false negatives just mean that some spam mails are classified as legitimate and moved to inbox. False positive mean that legitimate emails that get mistakenly identified as spam and moved to spam folder or discarded. For most users, missing legitimate email is an order of magnitude worse than receiving spam. The false positive rate of each classifier also considered to measure its performance.

The performance of the classifiers are summarized in Table II and shown in Fig.2 and Fig.3.

TABLE II
COMPARATIVE RESULTS OF THE CLASSIFIERS

Evaluation Criteria	**Naïve Bayes**	**J48**	**MLP**
Training time (secs)	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

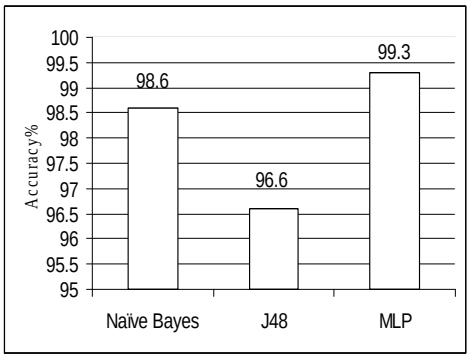


Fig. 1 Classification Accuracy

The performance of the three models was evaluated based on the three criteria, the prediction accuracy, learning time and false positive rate. Multilayer perceptron predicts better than other algorithms.

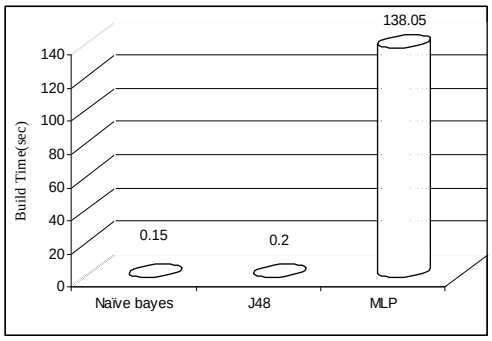


Fig. 1 Learning Time of the Models

Multilayer perceptron, the neural network classifier consumes more time to build the model. The naivebayes, the probabilistic classifier and decision tree model tends to learn more rapidly for the given data set.

VI. CONCLUSION

Although there are many email spam filtering tools exists in the world, due to the existence of spammers and adoption of new techniques, email spam filtering becomes a challenging problem to the researchers. In our work, we generated spam and legitimate message corpus from the latest mails and employed machine learning techniques to build the model. The performance of the model is evaluated using 10-fold cross validation and observed that Multilayer Perceptron classifier out performs other classifiers and the false positive rate also very low compared to other alogorithms. Email spam filters using this approach can be adopted either at mailserver or at mail client side to reduce the amount of spam messages and to reduce the risk of productivity loss, bandwidth and storage usage.

REFERENCES

[1] Ahmed Khorsi, "An Overview of Content-based Spam Filtering Techniques", *Informatica*, vol. 31, no. 3, October 2007, pp 269-277.

[2] Alistair McDonald, "SpamAssassin: A Practical Guide to Integration and Configuration", Ist Edition, Packt publishers, 2004.

[3] Ian H. Witten, Eibe Frank, "Data Mining – Practical Mahine Learning Tools and Techniques," 2nd Edition, Elsevier, 2005.