Abstract

The Internet E-mail infrastructure has become famous and widely used means of communication for personal, business and academic purposes just because it is fast, cheap and very efficient. People are using this E-mail infrastructure for their day-to-day work. The users are receiving many unwanted E-mails from the unknown senders. These unwanted E-mails are called as Spam E-mails.

The Spam in the context of E-mail is defined as, Unsolicited Commercial E-mail (UCE) or Unsolicited Bulk E-mail (UBE). The Spam is generally understood to mean the repeated mass mailing of UCE or UBE by a sender who disguises or forges or hide his real identity.

These Spam E-mails consumes valuable resources which includes network, bandwidth, data storage space. It reduces the productivity of employees as well as productivity of the business organisations.

Nowadays Denial of Service attack (DoS), which is also executed in distributed fashion called as DDoS and Phishing attacks are increased in number. These Phishing E-mails rides on the back of Spam E-mails, so it is important to block Spam E-mails.

This thesis addresses to the problem of Spam E-mails faced by users of E-mail Infrastructure by proposing the Anti-Spam Framework. This proposed Anti-Spam Framework consists of legislative measures, behavioural measures and technological measures.

In the Legislative measure of this proposed Framework, some important recommendations are made in Indian context after carrying out the study of Anti-Spam laws implemented by different countries with different parameters. The parameters considered for this study are type of Subscription, scope of subscription, sender as well receiver and possible accusers. In India it is recommended that, there is need to have separate Anti-Spam law. After implementing Anti-Spam law, wide publicity should be given to it so that, the users of E-mail will be aware of this laws and the provision made for punishments under this law. It is recommended that, there should be a single (possible online) reporting mechanism using which users can report the complaint of Spamming. Also using this mechanism, users should be able to upload the sample of Spam E-mail received by him/her. This reporting mechanism should become a data or sample collection tool. This sample will be given to the behavioural measure to study the pattern of Spam E-mail.
In the Behavioural measure of this proposed Framework, the study of E-mail delivery pattern is carried out along with structure of an E-mail. In this measure, the content analysis of header and body part of an E-mail is carried. During this content analysis some important observations are carried out along with pattern or combination of words are found out which are used by Spammer to send Spam E-mail. In this proposed Framework it is recommended that, the sample collected at reporting mechanism will be studied in behavioural measure section, to find out the behavioural pattern of Spam E-mail from it. This behavioural pattern is then given as input to technological measure to suggest the technological solution to filter or block Spam E-mails.

In Technological measure, the Origin based Filters (including White-list, Black-list and Challenge Response System) and Content based Filters (including ML based and Semantic Similarity based Classifier) are implemented. The empirical analysis of these techniques are carried out after applying the process of Feature Extraction on standard E-mail datasets such as Enron, LingSpam, PU123A, Spambase and PEM. The results obtained after these techniques are compared with the evaluation metrics such as Spam recall, Spam precision, accuracy, false positive and false negative. After carrying out comparative analysis, the technological solution to filter or block Spam E-mail is proposed which is combination of Origin based Filters and Content based Filters. This combination is adaptive in nature.

It is found that, the performance of Black-list and White-list is accurate for blocking Spam E-mails based on origin. The drawbacks of Challenge Response System are overcome by suggesting the architecture at sender as well as receivers side. So, these are combine together to form Origin based Filters.

The Content based Filter consist of ML based Classifier and Semantic Similarity with edge based approach Classifier. It is found that, in Content based Filter, Naive Bayesian Classifier, Rough Set with Genetic Algorithm and SVM with polynomial kernel outperforms. Also the performance of Semantic Similarity with edge based classifier is promising. All techniques are combined together for suggesting the Content based Filter. This proposed technological solution, consists of the combination of Origin based Filters and Content based Filter is adaptive in nature.

Thus, the complete Anti-Spam Framework is proposed for effective Spam management.