Chapter-1

Introduction

Nowadays, Internet E-mail infrastructure has became a medium 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.

Spam in the context of E-mail is defined as Unsolicited Bulk E-mail (UBE) or Unsolicited Commercial E-mail (UCE). Spam is generally understood to mean the repeated mass mailing of unsolicited commercial messages by a sender who disguises or forges or hide his real identity. Flooding Internet with many copies of the same Spam E-mail is known as ‘Spamming’. The people, who send Spam E-mails, are called as a ‘Spammers’. The Spammers can send thousands or millions of Spam E-mails at negligible cost [Pantel and Lin, 1998], [Schneider, 2003].

1.1 Types of Spam

The Spam E-mail today comes in different variety which is more harmful and may possess destructive characteristics. Spam is a form of abuse of the Simple Mail Transfer Protocol (SMTP). Most Spam making tools exploits the security holes of SMTP. They do this by forging E-mail headers, disguising sender addresses and hiding the sending system such that, it becomes difficult to identify the true sender. The users get Spam not only through E-mails but even while visiting a blog, chatting with a friend, browsing on community forum and also on mobile phones.

These Spam E-mails have lot of forms like, some of these contains advertisements, others provide winning notifications, and sometimes gets messages with executable files, which finally emerge as malicious codes, such as viruses and Trojan horses which can affect the recipients computers.

1.2 Problem caused by Spam

The design principles of the E-mail infrastructure, which were originally intended to provide simplicity and flexibility, have become ambivalent characteristics. The recipient pays a considerable price for receiving these unwanted Spam E-mails. The Spam E-mails causes the following economic harms:
A) Loss of Productivity :- When an employee receives Spam E-mails he/she spend time on opening, reading and classifying these E-mails as Spam and then spend time to delete them.

B) Download Cost: - It costs real money for the receivers to download their E-mails. Since, many receivers are still pay money for the time duration during which the content of mailbox are transferred from ISP to their computer.

C) Harm through malicious code:- Many Spam E-mails contain malicious code such as Viruses, Trojan horses, worms, spyware, adware and key logger. The economic harm that results from the execution of malicious software has not yet quantified.

D) Harm through frauds: - Frauds including phishing is an indirect harm that, Spam E-mail can cause, which has increased in number.

E) Decreases Bandwidth: - The bandwidth is precious resource in a corporate network. Spam E-mail essentially eats up lot of bandwidth. Spam also consumes other valuable computing resources. Smaller organizations working with minimal bandwidth are especially feeling the increasing strain that, Spam E-mail is putting on their network.

This in turn reduces the productivity of business organizations as well as of employees. Thus, Spam dissipated employee time, burdens on E-mail servers with a heavy processing load, eats up the disk space on client machines and degrades the overall network performances [Garcia et.al,2004] [Metsis et.al, 2006].

It is a common mistake to believe that, Spam E-mails are sent for marketing or money-making schemes. Some viruses and worms also send out Spam E-mail in order to infect the recipients' computers and carry out Denial of Service Attack (DoS). Nowadays, it is executed in distributed fashion called as Distributed Denial of Service Attack (DDoS)

Recently, it is found that, the incidents of phishing attacks are increased. The E-mail which carry phishing attack is called as ‘Phishing E-mail’. The Phishers execute phishing attack by hacking important information of users such as E-mail account information, Bank details,

Debit/Credit card numbers with PIN. These Phishing E-mails rides on the back of Spam E-mails. So, it is important to block Spam E-mails.
In this context many Spam filtering techniques are deployed at MTA. But majority of it do the mis-classification. Some legitimate E-mails are misclassified as Spam and vice-versa. So it is need of time to block the Spam E-mails only.

1.3 Problem Definition, Objectives and Contribution

1.3.1 Problem Definition
To develop effective Anti-Spam Framework for Spam management, this will keep Internet E-mail infrastructure alive as a reliable, cost-effective and flexible service.

1.3.2 Objectives
- Analysis of legislative and behavioural measures in Spam management.
- To propose an effective Spam classification scheme/s in technological Anti-Spam measures.
- To analyze effectiveness of Anti-Spam measures relative to the identified Spamming options.
- To develop a complete infrastructure Framework for Spam management

1.3.3 Contribution
- The problem of Spam E-mail is studied with different measures which include legislative measures, behavioral measure and technological measures. Since, legislative measures and behavioral measures are related to different stakeholders, the research is focused on technological measures after studying legislative and behavioral measures.
- After carrying out the study of legislative measure, it is suggested that, in India it is need to have separate Anti-Spam legislation. There should be an online reporting mechanism for reporting the incidents of Spam.
- The online reporting mechanism would serve as a data collection tool which would be useful for training the Content based Filters.
- In the study of behavioral measures the content analysis of Spam E-mail is carried out and some important observations are made which are utilized to propose technological solution.
- The Challenge Response System has two problems one is deadlock and another it blocks the automated newsletter. To overcome these two problems, solution is
proposed with the combination of Blacklisting and White listing which can be deployed at sender’s Mail Transfer Agent (MTA) and at receiver’s MTA.

- The important features of Spam E-mails and Ham E-mails are extracted from standard datasets such as Enron, Spambase, LingSpam and PU123A. In order to work with real life E-mails some Spam and legitimate E-mails are collected from personal E-mails received at ‘ajaysurwade@gmail.com’ which is referred as Personal E-mail Messages (PEM).

- The Content based Filter with Machine Learning (ML) based classifier is implemented. The empirical analysis of Spam E-mail is carried out by using classifiers such as Decision Tree, Rough Sets using various rule generation methods (Genetic Algorithm, Linear Algorithm, Covering Algorithm and Exhaustive Algorithm), k-Nearest Neighbor, and Support Vector Machine classifiers with various kernel functions (Linear, Multi-Layer Perceptrons, Quadratic Functions and Radial Basis Function).

- The Content based Filter using semantic similarity with edge based approach classifier is implemented which finds the semantic similarity between the words of E-mail to be tested with known Spam and Ham words. The decision is made depending on the shortest path length between the words. If the path length between the words of E-mail to be tested and known Spam words is shortest, then E-mail is classified as Spam E-mail. Similarly, if the path length between the words of E-mail to be tested and known Ham words is shortest, then E-mail is classified as Ham E-mail. The path similarity method is used on ‘WordNet’ dictionary to find semantic similarity between the words.

- The combination of Origin based Filter along with Content based Filter is implemented which is adaptive in nature.

- Finally, a complete Anti-Spam Framework is proposed for effective Spam management.

1.4 Organisation of Thesis

The structure of this thesis is as follows

- **Chapter-2: Literature Survey.** This chapter describes the prior and important work done in the area of blocking Spam E-mails. The intention of this survey is to provide readers with up to date contribution made by other researchers in the area of Spam filtering.
• **Chapter-3: Proposed Anti-Spam Framework.** This chapter gives an overview of proposed Anti-Spam Framework to address the problem of Spam E-mails. The framework consists of three important components such as legislative measures, behavioral measures and technological measures. It describes the system architecture of proposed technological solution along with system architecture of Origin based Filter and Content based Filter.

• **Chapter-4: Legislative Measures.** This chapter describes the study of different legislative measures applied all over the world to fight against the battle of Spam E-mails. During this study of anti-Spam legislative, the parameters such as type of subscription, scope of subscription, sender and receiver as well as possible accuser are considered. After studying these Anti-Spam legislations, some important recommendations are made in the context of India.

• **Chapter-5: Behavioral Measures.** This chapter is based on the observation carried out on Spam E-mails which are received daily. The Behavioural measures are important which can be applied complimentarily with legislative and technological measures to fight against the Spam. This chapter describes the content analysis of Spam E-mails which includes headers and body part of an E-mail. It describes study of E-mail delivery pattern and structure of an E-mail. Important observations are listed in summary and conclusion.

• **Chapter-6: Technological Measures.** This chapter has proposed technological solution to block the Spam E-mails. It consists of combination of Origin based Filters with Content based Filters which is adaptive in nature. The Origin based Filter consists of White-list and Black-list. Also, the Content based Filter is combination of Machine Learning (ML) based Classifier and Semantic Similarity (SS) with Edge based Classifier. The important features are extracted from the collection of standard datasets like Enron, Spambase, LingSpam, PU123A and PEM. This chapter contains the empirical analysis of Content based Filters is carried out. The results are discussed in terms of accuracy, Spam precision, Spam recall, false positive and false negative. It also describes the implementation of Semantic Similarity with Edge based approach Classifier.
Chapter-1: Introduction

- **Chapter-7: Result and Discussion.** In this chapter the findings of research work carried out, are summarized. These results are compared with the previous research work and comparative discussion is carried out.

- **Chapter-8: Conclusion and Future Directions.** In this chapter the overall conclusion is drawn on Legislative, Behavioral and Technological measures and finally, the direction for future work is addressed.