Chapter-3

Proposed Anti-Spam Framework

Every day, the Spammer finds out new technique to send Spam E-mail. The Anti-Spam solutions are proposed after studying the current pattern of Spamming. Next day, Spammer, finds another techniques which bypass the Anti-Spam solutions. These Anti-Spam solutions are again studied by intelligent spammer and they find out methods to bypass it again and send another pattern of Spam E-mails. Thus, it is game of cat and mouse which will continue.

In order to provide effective Anti-Spam solution, it is necessary to study the problem of Spam E-mails using different measures such as legislative, behavioural and technological measures. The proposed Anti-Spam framework described in the next section, addresses these measures, which is followed by the description of standard datasets used for carrying out the experiments along with collection of Personal E-mail Messages (PEM). The last section describes the Evaluation Metrics which are used for comparative analysis.

3.1 Introduction

The proposed Anti-Spam Framework consists of three building blocks such as legislative measure, behavioural measure and technological measures. The block diagram of proposed framework is as shown in Figure-3.1.

These measures are described in the following sections separately.

3.2 Legislative Measures

The legislative measure is important measure to fight against Spam. The study of Anti-Spam legislation is carried out. This study mainly focused on the different laws passed by the different countries to address the problem of Spam E-mail considering the parameters such as type of subscription, scope of subscription, sender
as well as receiver and possible accuser. There are still many countries which have no or low effective Anti-Spam laws and where Spamming is tolerated. The detail description and finding are explained in Chapter No.4. The important recommendations are made in the Indian context which includes the need of online mechanism where user can register complaint of incidents of spamming and can submit the sample of Spam E-mails. The reporting mechanisms would become data collection tool, which can be useful for Content based Filters to learn or understand the current pattern of Spam E-mails. This online reporting mechanism is an important outcome of legislative measure, which would be useful data or sample collection tool and this data or sample will be given as input to the behavioural measure for further content analysis which is described in Chapter-5.

3.3 Behavioural Measures

Another important section of the proposed framework is behavioural measure. The study of behaviour of Spam E-mail is useful to find current pattern of it using which the Spam words can be collected and Content based Filters as well as Origin based Filters are updated. The samples collected using online reporting mechanism suggested in legislative measure, would be analysed for carrying out the behavioural study of Spam E-mail. This will help technological measure to study the current pattern used by Spammers for sending Spam E-mails with the sample of it. The behavioural study becomes important as it becomes useful to propose the technological solution to block the Spam E-mail. The study of behavioural measure along with some observation of Spam E-mail is briefly explained in the Chapter-5. The observations made after studying the samples and patterns of Spam E-mails are used for suggesting the technological solution to block Spam E-mails.

3.4 Technological Measures

The important output of legislative measure is the collection of Spam samples which is uploaded by users who has received Spam E-mails. This sample is studied in the behavioural measure to find out the pattern of Spam E-mails. This pattern is used in technological measure to propose an effective technological solution. The system
architecture of proposed technological solution to fight against the Spam E-mail is shown in Figure-3.2.

![System Architecture for Technological Solution](image)

**Figure-3.2: System Architecture for Technological Solution**

This system architecture includes important building blocks such as process of Feature Extraction along with combination of Origin based Filters and Content based Filters which are adaptive in nature.

The incoming E-mail which needs to be classified as Spam or Ham E-mail is given as input to the Feature Extraction process where the features of E-mail such as IP or domain names as well the words are extracted. These features are given as input to Origin based Filter and Content based Filter. Each of these filter, then classifies the E-mail as Spam or Ham. The final decision is made using the Equation-6.16 to Equation-6.19 as explained in Chapter- 6, Section-6.6.1. This combination of Origin based Filter and Content based Filter is adaptive in nature as explained in Chapter No.6, Section-6.6.3. Thus, the E-mail is classified as Spam or Ham.

### 3.4.1 Origin based Filters

The Origin based Filter consist of two important lists such as White-list and Blacklist. These White-list and Black-list are described in Chapter No.6, Section 6.2.1 and 6.2.2 respectively.

The IP or domain names are given as input to Origin based Filters for testing whether it is white-listed or blacklisted. The Origin based Filters finds, whether the inputted IP or domain names are White-listed or Black-listed. Thus, decision is made based on the result returned by Origin based Filter.

If Origin based Filter finds the IP or domain name in White-list then, it is classified as Ham E-mail. If Origin based Filter finds the IP or domain name in Blacklisted then, it
is classified as Spam E-mails. The system architecture of Origin based Filter is shown in Figure-3.3.

3.4.2 Content based filters

The collection of words extracted from the incoming E-mail to be tested as Spam or Ham is given as input to Content based Filters. The Content based Filter consists of two classifiers such as Machine Learning (ML) based Classifier and Semantic Similarity (SS) with Edge based approach Classifier. Both of the classifier classifies the incoming E-mail as Spam or Ham using Equation-6.18 and Equation-6.19 as explained in Chapter No. 6, Section-6.6.2. The system architecture of Content based Filter is as shown in Figure-3.4.
Initially, the ML based classifier including Naive Bayesian Classifier, Decision Tree, Rough Set using various rule generation methods (Genetic Algorithm, Linear algorithm, Covering Algorithm, Exhaustive Algorithm), k-Nearest Neighbor, and Support Vector Machine classifiers with various kernel functions (Linear, Multi Layer Perceptrons, Quadratic Functions, Radial Basis Function) are implemented and tested on the standard datasets.

After carrying out empirical analysis of the above techniques it is found that, Naive Bayesian Classifier, SVM with Polynomial Kernel and Rough Set with Genetic Algorithm classifiers have produced promising results.

Another classifier of Content based Filter using Semantic Similarity with Edge based approach classifier is implemented which uses the path similarity method to find semantic similarity between the words to test and known Spam words and Ham words. The decision is made depending on the shortest path length between the words. If the path length between the words of Email to be tested and known Spam words is shortest, then Email is classified as Spam Email. Similarly, if the path length between the words of Email to be tested and known Ham words is shortest, then Email is classified as Ham Email, which is briefly explained in Chapter 6, Section 6.5.

Thus, an Email is classified as Spam or Ham based on the decision made by combination of Origin based Filter and Content based Filter using Equation 6.18 and 6.19. The next section describes, the standard dataset used by the classifiers for the experiments.

3.5 Standard Datasets

In this section, the structure of standard datasets such as Enron, LingSpam, PU123A, Spambase including Personal Email Messages are discussed.

3.5.1 Enron Dataset

The standard dataset Enron consists of six sub-folders introduced by [Klimt and Yang, 2004]. This dataset was collected and prepared by Cognitive Assistant that Learns and Organizes (CALO) Project [Enron, 2012]. This data was originally made public and posted on the web, by the Federal Energy Regulatory Commission (FERC) during its investigation. Table-3.1 gives at glance information of Enron dataset.
Table-3.1: Enron dataset

<table>
<thead>
<tr>
<th>Name of Folder</th>
<th>No. of files</th>
<th>Ham files</th>
<th>Spam files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enron-1</td>
<td>385</td>
<td>‘farmer.ham’</td>
<td>-----</td>
</tr>
<tr>
<td>Enron-2</td>
<td>5858</td>
<td>‘kaminski.ham’</td>
<td>‘SA+HP.spam’</td>
</tr>
<tr>
<td>Enron-3</td>
<td>5513</td>
<td>‘kitchen.ham’</td>
<td>‘BG.spam’</td>
</tr>
<tr>
<td>Enron-4</td>
<td>6001</td>
<td>‘williams.ham’</td>
<td>‘GP.spam’</td>
</tr>
<tr>
<td>Enron-5</td>
<td>5176</td>
<td>‘beck.ham’</td>
<td>‘SA+HP.spam’</td>
</tr>
<tr>
<td>Enron-6</td>
<td>6001</td>
<td>‘lokay.ham’</td>
<td>‘BG.spam’</td>
</tr>
</tbody>
</table>

3.5.2 LingSpam Dataset

The standard dataset LingSpam described in [Androutsopoulos et.al, 2000] is a collection of both Spam and Ham messages which contain one message in each file. Each message consists of linguistic Ham and Spam words. The files with name ‘spmsg*.txt’ are Spam files while, file name with ‘msg.txt’ are Ham files. The LingSpam consists of 481 Spam messages and 2412 Ham messages these are received from the volunteer providers [LingSpam, 2012]. The LingSpam dataset, in spite of its imperfect nature, is the best dataset which can be used for testing the Spam filters [Androutsopoulos et.al, 2004] [Wenbin Li et.al, 2007].

3.5.3 PU123A Dataset

Another standard dataset PU123A contains both Spam and Ham messages, one message in each file. Files name with the format ‘*spmsg*.txt’ are Spam messages while, files with format ‘*legit*.txt’ are Ham or Legitimate messages [PU123A, 2012]. These dataset are encoded with numbers.

3.5.4 Spambase Dataset

The Spambase dataset is created by Mark Hopkins, Erik Reeb, George Forman, Jaap Suermond at Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304. It consists of total 58 attributes. Out of which 57 attributes are word attributes and one attribute is of class ‘Spam’ or ‘Ham’ [Spambase, 2012].

3.5.5 Personal E-mail Messages Dataset

In order to test real life E-mails some Personal E-mails are collected as set of Ham E-mail and Spam E-mails received at the E-mail address ‘ajaysurwade@gmail.com’. These Ham and Spam E-mails are classified as Ham and
Spam by the techniques used at the domain name, ‘www.gmail.com’ at level of MTA or MDA. The Ham folder contains the personal Ham files which reached to the Inbox of the E-mail address ‘ajaysurwade@gmail.com’. The Spam E-mail which reached to Spam folder of this E-mail address is stored in the Spam folder of the dataset. In order to keep originality and all concern information of an E-mail at one place each E-mail has been saved in a separate text file.

3.6 Evaluation Metric

The results obtained by the classifier are converted in the form of Confusion Matrix. The Table-3.2 shows the format of Confusion matrix used for the Spam E-mail classification.

The Evaluation Metric such as accuracy, Spam precision, Spam recall, false positive and false negative are the most important performance Evaluation Metrics which are used by [Turtle et.al, 2004], [Subramaniam et.al, 2010], [Md. Islam et.al, 2009], [Awad and ELseauofi, 2011] the researchers in the Spam filtering.

**Table-3.2: Confusion Matrix**

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ham</td>
<td>Spam</td>
<td></td>
</tr>
<tr>
<td>Ham</td>
<td>True Positive</td>
<td>False Positive</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(TP)</td>
<td>(FP)</td>
<td></td>
</tr>
<tr>
<td>Spam</td>
<td>False Negative</td>
<td>True Negative</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(FN)</td>
<td>(TN)</td>
<td></td>
</tr>
</tbody>
</table>

**Spam Precision**

Spam Precision (SP) represents the ratio between the numbers of correctly classified Spam to the number of all messages marked as Spam. The Spam precision is calculated by using the Equation-3.1.

\[
SP = \frac{TN}{TN+FP} \times 100\% 
\]  

(3.1)

**Spam Recall**

Spam Recall (SR) indicates the number of correctly classified Spam to the number of correctly classified Spam and number of misclassified as Ham. The Spam recall is calculated by using the Equation-3.2.

\[
SR = \frac{TN}{TN+FN} \times 100\% 
\]  

(3.2)
Chapter-3: Proposed Anti-Spam Framework

Accuracy
Accuracy (A) represents the ratio between the number of correctly classified Spam and Ham E-mails to the total E-mails used for testing that is all E-mails that are correctly classified by the classifier. The Accuracy is calculated using Equation-3.3.

\[ A = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \]  

(3.3)

False Positive
False Positive (FP) also called as False Spam is defined as the ratio of Ham E-mail is classified as Spam to the total Ham. It is misclassification as a Spam. The false positive is calculated by Equation-3.4.

\[ FP = \frac{FP}{TP+FP} \times 100 \% \]  

(3.4)

False Negative
False Negative (FN) also called as False Ham is defined as the ratio of Spam E-mail is classified Ham to the ratio of total Spam. It is misclassification as a Ham. The false negative is calculated by Equation-3.5.

\[ FN = \frac{FN}{FN+TN} \times 100 \% \]  

(3.5)

3.7 Conclusion
The Spammer finds out new technique every time to send Spam E-mail. The Anti-Spam framework is proposed in this chapter consists of legislative, behavioural and technological measures. The structure of standard dataset such as Enron, LingSpam, PU123A, and PEM is also discussed which are used to carry out the experiments. The Evaluation Metrics such as Spam recall, Spam precision, accuracy, false positive and false negative are discussed.