Email Filtering and Analysis Using Classification Algorithms

Akshay Iyer, Akanksha Pandey, Dipti Pamnani, Karmanya Pathak and Prof. Mrs. Jayshree Hajgude

IT Dept, VESIT
Chembur, Mumbai-74, India

IT Dept, VESIT
Chembur, Mumbai-74, India

IT Dept, VESIT
Chembur, Mumbai-74, India

IT Dept, VESIT
Chembur, Mumbai-74, India

Abstract

With the various developments that are taking place in the field of technology especially in the communication domain, there are a plethora of malpractices that are being practiced to hinder the users` chores. Most of these practices can be observed in the Email Account of a user. Any internet user with a registered email account has an inbox which consists of his emails. These mails include everything from personal messages to advertisements and entertainment promotional emails. All these emails are present in an unorganized manner. Also some mails which are being received by the user may contain harmful content which may prove to have severe consequences (Normally Termed as Spam). The motivation behind this paper is to try and find a solution to free the users of internet from this issue pertaining to their emails account and emails. With this idea in mind, we present this report on Email Filtering and Analysis using Classification Algorithms.

Keywords: *Naive Bayes, C4.5, Spam, Non Spam, Email, Keyword.*

1. Introduction

Email Filtering is the process which is used in order to classify the emails into various categories on the basis of their content. We have developed a desktop application to serve this purpose and ease out the users` tedious task of segregation of his emails. Since, we have developed a desktop application which is password protected for every single user of the application, we take the liberty of accessing the content of the emails fetched from the user`s email account. The application fetches the emails from a user's id, and stores it on a server. It then classifies the emails into spam and non-spam using classification algorithms of Naïve Bayes and C4.5.

Also, the application classifies the emails into user defined categories on the basis of the keyword entered by the user. The user can also send, forward and reply to a particular email. There is also a lot of historical spam analysis done by the application on the basis of the content downloaded by the user. The user can access, read, store and copy the contents of his email.

Further, the two most commonly used algorithms namely Naïve Bayes and C.45, have been compared and their performance characteristics have been visually represented. Thus, our application not only classifies emails as per user requirements but also analyses the classification process.

1.1 What is Email Filtering?

Email Filtering, in the context of our application, refers to the classification of an account's emails based on two types of emails (unless keywords specified by the user):

1. Spam and
2. Non-Spam.

The user first registers with the application by selecting an available username and setting a password for the account. He then logs in to his account using the registered id and the corresponding valid password. Upon logging in, the user's mails are fetched in the database and are classified into spam and non-spam. The user can also create custom labels which are classified using keywords provided by the user. Also, he can browse for the unread and read emails. This makes the mail service easy and user friendly.

A basic task in email filtering is to mine the data from an email and to classify it into the different categories using data mining classification algorithms.

Email Filtering involves spam filtering, generalized filtering and segregation and filtering of inbound emails. Spam mails are filtered since they are not important to most of the users. Generalized filtering and segregation of emails is segregation of the mails into different categories as specified by the user using custom labels.

Companies filter outbound emails so that sensitive data regarding the working of the company does not leak intentionally or accidentally by emails.

To summarize email filtering

- Segregates inbound mails into different categories.
- Filters outbound mails so as not to leak sensitive information.

1.2 Motivation for this domain

With the increase in the internet users, communication and transfer of files and data through different methods over the internet has increased drastically. In such times, it is difficult to know what kinds of emails are entering your organization or system.

Most of the present filtering techniques are unable to handle frequent changing scenario of emails adopted by the senders over the time.

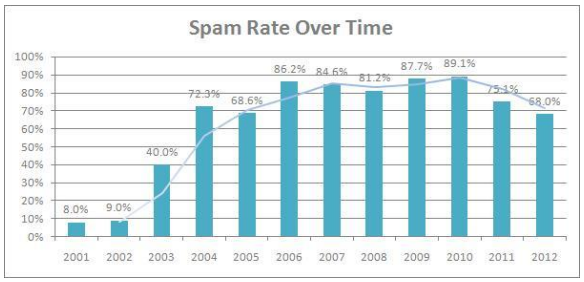

Fig. 3 Spam Rate over Time

In absolute numbers, the average number of spam mails sent per day increased from 2.4 billion in 2002 to 300 billion in 2010.

Google Today announced it has made security improvements to Gmail to further protect users' emails from snooping. Gmail now uses an encrypted HTTPS connection when users check or send emails and encrypts all messages moving internally on Google's servers.

With the advent of growth in technology, desktop based email applications are more increasingly used. Outlook express has changed the way the world communicates with the help of emails.

1.3 Different areas of Application

1.3.1 Spam Filtering

With the advent of Internet, the number of spam mails has increased manifold. A spam filter is a program that is used to detect unsolicited and unwanted emails and prevent those messages from getting to a user's personal inbox space. Like other types of filtering programs, a spam filter looks for certain criteria on which it bases judgments.

1.3.2 Filtering and Segregation of E-mails

Email filtering is the processing of emails to organize the emails according to specified criteria. Most often this refers to the automatic processing of incoming messages, but the term also applies to the intervention of human intelligence in addition to anti-spam techniques, and to outgoing emails. Filtering mails based on classes like spam, travel, social and a country-based classification of official mails for ease of access to emails from specific sub-branches would

help make the email service more efficient in terms of accessibility and user-friendliness.

1.3.3 Inbound and Outbound Filtering of E-mails

Email filters can operate on inbound and outbound email traffic. Inbound email filtering involves scanning messages from the Internet addressed to users protected by the filtering system or for lawful interception. Outbound email filtering involves the reverse – scanning email messages from local users before any potentially harmful messages can be delivered to others on the Internet. One method of outbound email filtering that is commonly used by Internet Service Providers (ISPs) is transparent SMTP proxy, in which email traffic is intercepted and filtered via a transparent proxy within the network. Outbound filtering can also take place in an email server. Many corporations employ data leak prevention technology in their outbound mail servers to prevent the leakage of sensitive information via email.

1.4 Issues Faced

1.4.1 Avoidance of vocabulary treated as Spam by Spammers

The subject and body content are chosen carefully by spammers. Being aware of terms, text processing rules of a filter, etc. helps the spammers to use alternate words still serving the same purpose yet not falling prey to the filter. This helps them to pass the filter and the mail is treated as a non spam mail and the spammer succeeds.

1.4.2 The Double Opt-In problem

One of the main problems faced by spammers is to gain access and explicit permission to mail any particular user. An efficient solution found out by the clan is the Double Opt-In method.
It works in the following manner:

1. The user enters his email address into an online form.
2. They receive a confirmation link.

On clicking the conformation link the spammer gets explicit permission to send mails to the user. These mails, though actually spam, are then treated as normal and non-spam mails.

1.4.3 The Encrypted E-Mail Problem

The Encrypted E-Mail Problem is one of the most important problems which are being faced by various E-Mail Client Applications. Most of the bank transactions which are being performed by various banks and corporate companies are sent in an encrypted format to the concerned user. This is done in order to ensure security. Many mails which are sent by many telecom and multinational companies concerning any payment or any transfer of money are also done in the encrypted format. The message which is viewed in the user inbox is not actually the email which has been revived by it. It is encrypted using some encryption key which can be retrieved by some user credentials, such as the user bank account number, his password. Thus, it is extremely difficult to bring about classification of mails in this format. Recently, Gmail had announced that, it has taken a step forward in correct classification of encrypted mails, which is soon to be implemented by them.

2. Analysis

2.1 The C4.5 Algorithm

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set

$$S = s_1, s_2, \ldots \quad (1)$$

of already classified samples. Each sample s_i consists of a p-dimensional vector

$$(x_{1,i}, x_{2,i}, \ldots, x_{p,i}) \quad (2),$$

Where they x_j represent attributes or features of the sample, as well as the class in which s_i falls.
At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. Thus, the C4.5 algorithm then recourses on the smaller sub lists.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

Pseudo code:

In pseudo code, the general algorithm for building decision trees is:

1. Check for base cases.
2. For each attribute a find the normalized information gain ratio from splitting on a.
3. Let a_best be the attribute with the highest normalized information gain.
4. Create a decision node that splits on a_best.
5. Recourse on the sub lists obtained by splitting on a_best, and add those nodes as children of node.

2.2 The Naïve Bayes Algorithm

A Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "Independent Feature Model". An overview of statistical classifiers is given in the article on pattern recognition.

In simple terms, a naive Bayes classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features.

For some types of probability models, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood; in other words, one can work with the Naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, Naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of Naive Bayes classifiers.

Probabilistic model:

Abstractly, the probability model for a classifier is a conditional model

$$p(C|F_1,\ldots,F_n) \tag{3}$$

over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables F_1 through F_n. The problem is that if the number of features n is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible.

Using Bayes' theorem, this can be written

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}. \tag{4}$$

In plain English, using Bayesian Probability terminology, the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}. \tag{5}$$

3. Implementation

3.1 Email Filtering

The Email Client window is the major window in the application. The major functionalities which are to be implemented are a part of the Email Client Window. The Email Client Window is entirely divided into 6 different parts, and each of these 6 parts is represented by 6 tabs which are present on the top of the Email Client Window.

All the functionalities can be accessed from the Email Client Window.

The entire Email Client Window is comprised of the following 6 tabs:

1. The welcome tab is the basic homepage where the user can view all the basic information, like how many mails have been downloaded, how many are unread, etc.

2. It is here that the user performs all the necessary operations, with respect to the client application. The user executes Naïve Bayes, and C4.5 classification algorithms, as well as can search for specific user defined keywords.
3. The user can view all his mails on the basis of the conditions that have been specified in this window, the message viewer helps the user read his mails, as per his preference.

4. The statistics window showcases graphical and historical analysis on the information that is made available to him from previously fetched data.

5. The user can send a message to another user, from the desktop application to a particular user's Email Account.

6. The connection window is the major window which takes all the login credentials and the required information from the user and stores it in the server.

3.2 Detailed Implementation

The Email Filtering Application has been implemented as a core Java Application. The front end of the application has been implemented using JFrame. The implementation of the email functionality in java has been brought about with the help of classes, functions, and interfaces provided by the 'JavaMail' package i.e. JavaMail API. The graphs in the application have been implemented using JFreeChart API.

The following are the various windows in our Email Application.

- The Connection Window
- The Welcome Tab
- The Main Page
- The Message Viewer
- The Statistics Window
- The Messaging Window
- The Message Dialog Box

3.2.1 The Connection Window

The signup credentials take information such as, the username, the password, and the name, surname, country, and mobile number of the user. The user also needs to provide the server with which he is going to be interacting, and the server which is going to be used by the user to perform message sending operations. As specified earlier, the two mail servers which are going to be accessed are the IMAP server, and the SMTP server. These servers are going to be used for message transport and access.

3.2.2 The Welcome Tab

It is a graphical display of how many mails the user has received. It further displays how many of these received emails have been read and how many are still unread. The red portion in the pie chart represents the total number of unread mails which the user is currently having in his mailbox. The refresh button allows the user to refresh his mailbox, so as to retrieve those emails which haven't been retrieved yet. This happens on the execution of the connect method which is executed by clicking on 'connect' from the connect dialog box.

3.2.3 The Main Page

The main page is the window where major classification operations are being performed. There are two algorithms that are being used, Naïve Bayes and C4.5. The classification is being performed using the training dataset which is imported and then various operations with respect to the dataset are performed by the user.

3.2.4 The Message Viewer

The message viewer gives the users a segregated view of the emails. The message viewer also gives the user a customized view of emails based on keywords specified. There are two additional buttons which have been provided; one is to store the particular file in a specific location which is defined by the user. The other feature is to copy all the message contents to the clipboard.

3.2.5 The Statistics Window

The statistics window is extremely useful in achieving historical analysis of mails, as to how many spam and non-spam emails have been received over the past few years.

Annual Statistics:

The annual statistics generate statistics from 2007 to 2017 and showcase how many emails have been received each
Year, how many of them are spam, and how many of them are non-spam.

Monthly Statistics:

The yearly statistics which are developed can be further simplified to a view that describes the number of spam and non-spam emails received each month during a particular year.

The Monthly Statistics can be viewed from the month of January and it continues till the month of December. All the months have been specified.

Weekly Statistics:

The monthly statistics which are developed can be further viewed on weekly basis.

The Weekly Statistics can be viewed in spans of 4 weeks of each month and this can be done for all months of a year. All the weeks have been specified.
Week 1: 1-7
Week 2: 8-14
Week 3: 15-21
Week 4: 22-31

Comparative Analysis:

This method shows a comparison between Naïve Bayes and C4.5 and tells the user, which algorithm has more efficiently and correctly classified emails as spam and non-spam.

3.2.6 The Messaging Tab

This tab helps the user to send emails, using the desktop application itself. The user can also select a particular message and forward that message to anyone. The user can also reply to emails which he has received. All these features have been implemented with the help of the Message Dialog box.

3.2.7 The Message Dialog Box

The message dialog box is the dialog box which has been provided to send a new email, reply to an email, or forward an email.

4. Screenshots

The following are some of the screenshots of the application, as well as some of the Modeling Diagrams.

4.1 The Connection Window

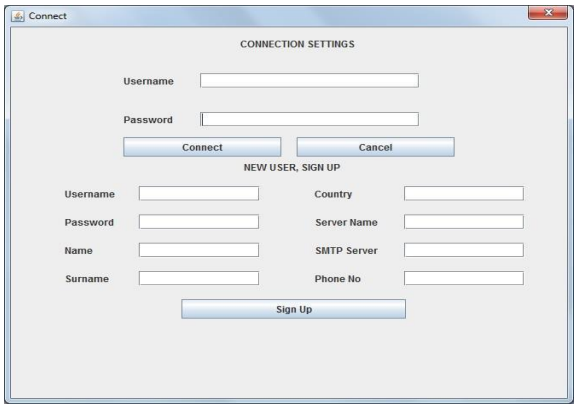

Fig. 2 SCREENSHOT 1

4.2 The Welcome Window

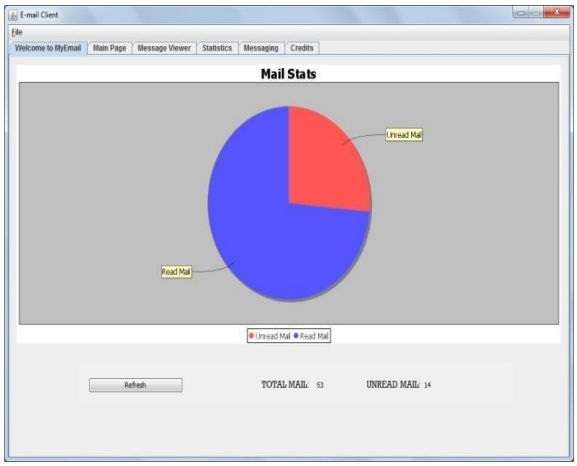

Fig 3. SCREENSHOT 2

4.3 The Main Page

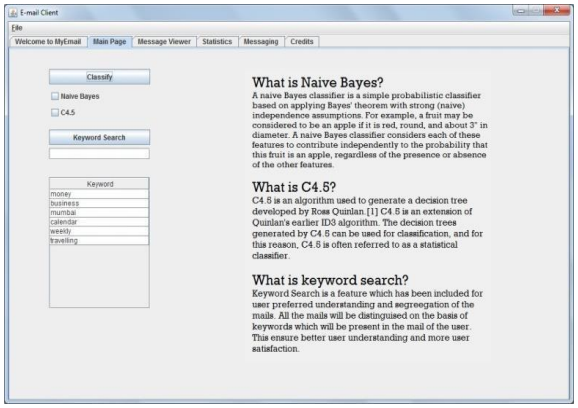

Fig. 4 SCREENSHOT 3

4.4 The Message Viewer

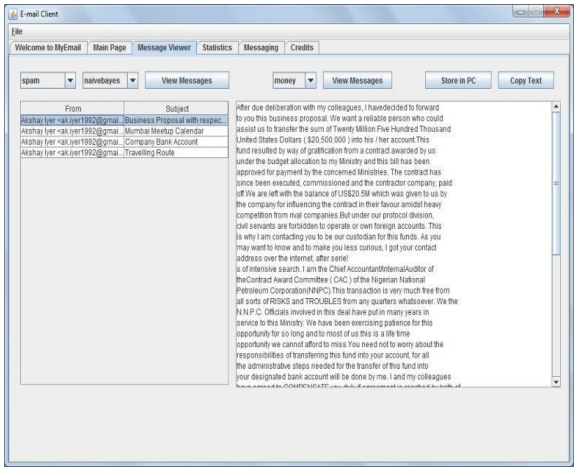

Fig. 5 SCREENSHOT 4

4.5 The Statistics Window

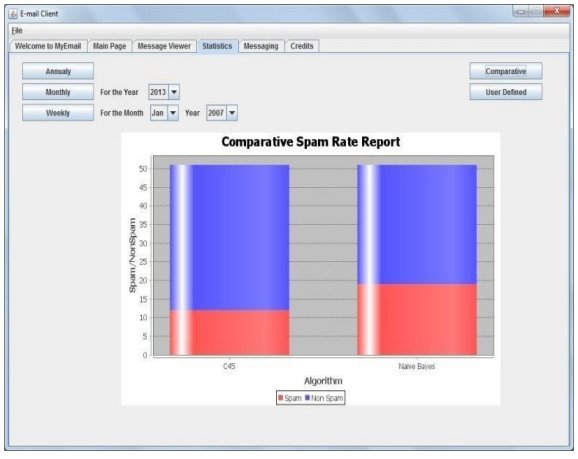

Fig. 6 SCREENSHOT 5

4.6 The Statistics Window – Annual, Monthly, Weekly

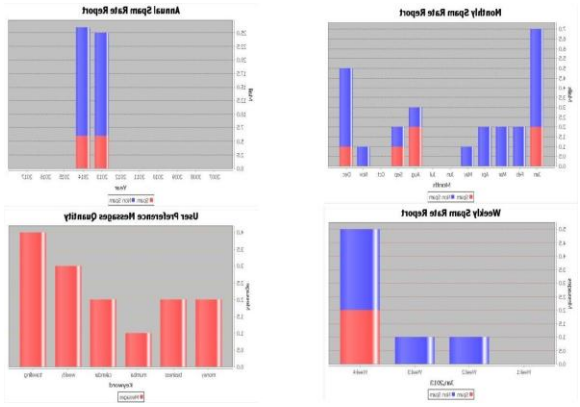

Fig. 7 SCREENSHOT 6

4.7 The Messaging Window

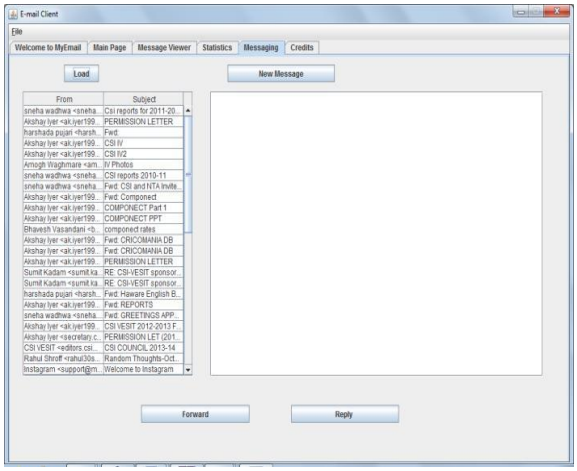

Fig. 8 SCREENSHOT 7

4.8 The Message Dialog Box

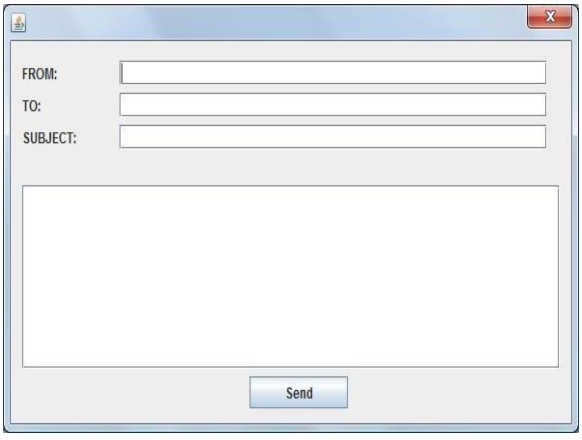

Fig. 9 SCREENSHOT 8

4.9 Reply Message Box

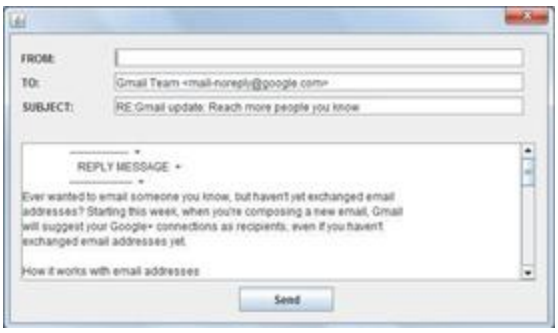

Fig. 10 SCREENSHOT 9

4.10 Forward Message Box

Fig. 11 SCREENSHOT 10

4.11 Activity Diagram

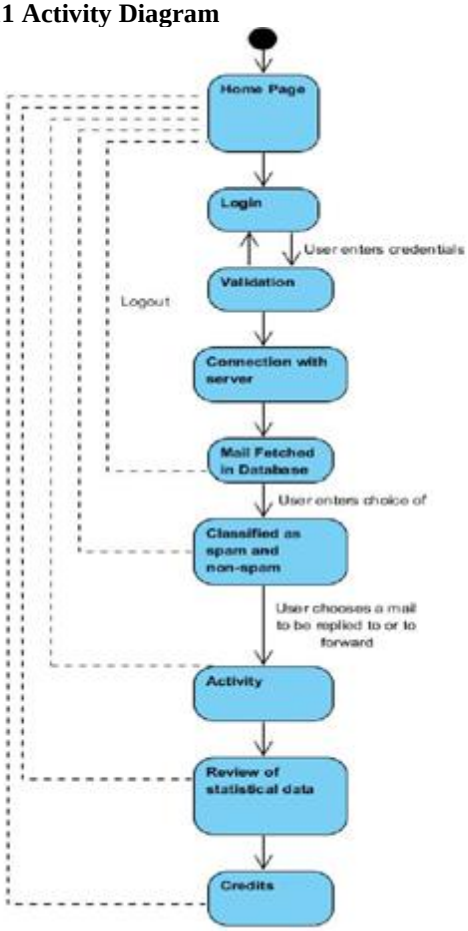

Fig. 12 SREENSHOT 11

4.12 Use - Case Diagram

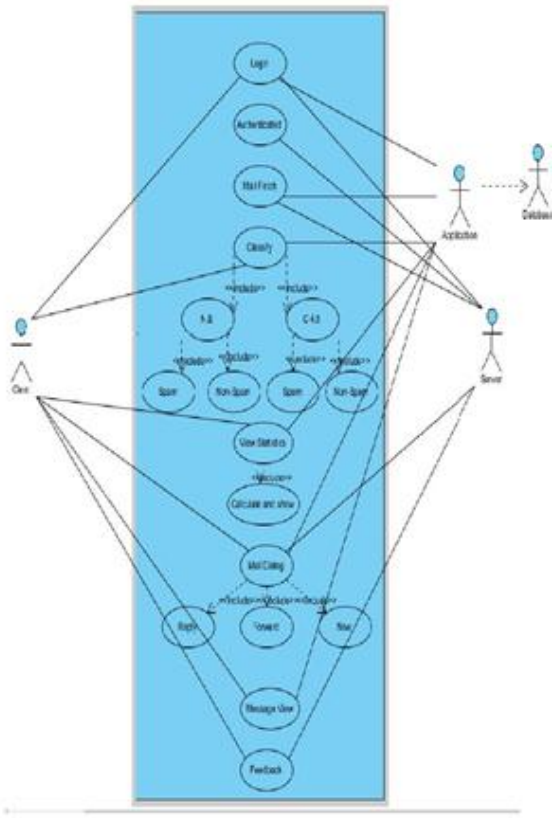

Fig. 13 SCREENSHOT 12

4.12 Class Diagram

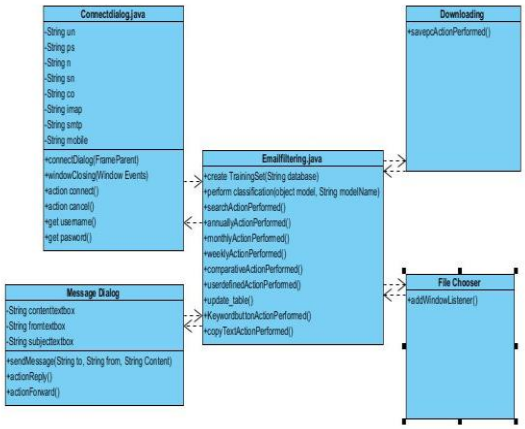

Fig. 14 SCREENSHOT 13

5. Conclusion

Considering the necessity of emails in an individual's life, the need of classifying the messages is of utmost importance and it is necessary to be achieved. With the employment of various spam filtering techniques and various classification algorithms, it has become extremely easy to classify the information into various categories. Thus, through our paper we tried to provide an insight into two of the commonly used classification approaches, their effectiveness and how classification and data mining approach can simplify the users` tasks and provide a better human interface.

6. Future Scope

6.1 Cloud Based Email Archiving System

The concept of cloud based email archiving is pretty simple. Broadly put, a service provider typically processes, manages and stores our business data in a hosted server and at a remote place either as a substitute or typically as an enhancement to our on premise infrastructure.

6.2 Encrypted message based E-Mail Classification

This is an application which will enable the user to fetch messages from the server and perform classification on the message only after decrypting the encrypted messages. On the basis of the information obtained, the application will decrypt the text obtained from the e-mail server and execute the classification algorithms. On the basis of the results obtained, the best solution will be selected amongst all the decrypted texts. If however, the algorithm fails to decrypt the text, then the message will be passed as non-encrypted text and further filtering according to the categories will take place.

6.3 An Android Based Application for accessing Emails

An android based application can be developed to enable the user to access his emails from any location. We could make use of the same server to access and store emails.

6.4 Location based Analysis of Spam Rate

Location based analysis of spam can prove to be a really helpful feature that can be implemented in the

future. It will help the user the know mails from which location are generally spam and identify a zone of spammers. This can be graphically displayed using Google Maps and Java map.

7. References

[1] Videos on Java Swing programming by 'Programming Knowledge` on www.youtube.com

[2] Sun Certified Java Programming-Kathy Sierra and Bert Bates.

[3] Information on the Naive Bayes Algorithm http://en.wikipedia.org/wiki/Naive_Bayes_classifier

[4] Information on the C4.5 Algorithm http://en.wikipedia.org/wiki/C4.5

[5] A Bayesian Approach to Filtering Junk Email http://robotics.stanford.edu/users/sahami/papers-dir/spam.pdf

[6] An Evaluation of Naive Bayesian Anti-Spam Filteringhttp://arxiv.org/pdf/cs/0006013.pdf

[7] Information on using JFreeChart http://www.jfree.org/jfreechart/samples.html

[8] Data Mining Concepts and Techniques- Jiawei Han, Micheline Kamber, Jian Pei.

Akshay Iyer

Akshay Iyer has graduated from Vivekanand Education Society's Institute of Technology in 2014. He is currently employed with Indus Valley Partners.

Akanksha Pandey

Akanksha Pandey has graduated from Vivekanand Education Society's Institute of Technology in 2014. She is currently employed with Accenture.

Dipti Pamnani

Dipti Pamnani has graduated from Vivekanand Education Society's Institute of Technology in 2014. She is currently employed with Accenture.

Karmanya Pathak

Karmanya Pathak has graduated from Vivekanand Education Society's Institute of Technology in 2014.