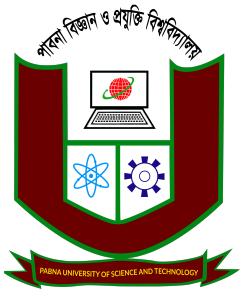
Spam Mail Filtering Based on Machine Learning and Analyzing User and Email Behavior



Department of Computer Science and Engineering

Pabna University of Science and Technology, Pabna-6600

Course Title: Thesis

Course Code: CSE 4100 and CSE 4200

A thesis has been submitted to the Department of Computer Science and Engineering for the partial fulfillment of the requirement of Bachelor of Science in Computer Science and Engineering

Submitted By:

Md Masud Rana, Roll Number:160142

Registration Number:101719, Session:2015-16

Supervised By:

Md. Shafiul Azam

Associate Professor, Department of Computer Science and Engineering

Pabna University of Science and Technology

January, 2022

DECLARATION

In accordance with rules and regulations of Pabna University of Science and Technology following declarations are made:

I hereby declare that this thesis has been done by me under the supervision of Md. Shafiul Azam, Associate Professor, Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna-6600.

I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for awarding of any degree and any material reproduced in this thesis has been properly acknowledged.

Signature of the Examinee

CERTIFICATE

I am pleased to certify that Md Masud Rana, Roll Number: 140142, Registration Number: 101719, Session: 2015-16 has performed a thesis work entitled "Spam Mail Filtering Based on Machine Learning and Analyzing User and Email Behavio" under my supervision for the requirement of the completion of course entitled 'Thesis'. So far as I concern this is an original thesis that has been carried out for one year in the Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna-6600, Bangladesh.

To the best of my knowledge, this paper has not been duplicated from any other paper or submitted to elsewhere prior submission to the department.

Md. Shafiul Azam

Associate Professor,

Department of Computer Science and Engineering

Pabna University of Science and Technology, Pabna-6600.

Bangladesh.

ACKNOWLEDGEMENT

First of all I would like to admit my gratefulness to the Almighty God for enabling me to perform this thesis successfully. I would like to express my deepest sense of gratitude to my honorable supervisor Md. Shafiul Azam, Associate Professor, Department of Computer Science and Engineering (CSE), Pabna University of Science Technology (PUST), for his scholastic supervision, valuable guidance, adequate encouragement and helpful discussion throughout the progress of this work. I am highly grateful to him for allowing me to pursuing this study under his supervision.

I am deeply thankful to honorable chairman,Dr.Md. Abdur Rahim, and all the respectable teachers of the Department of Computer Science and Engineering, Pabna University of Science Technology, Pabna-6600, Bangladesh, for their encouragement to my research work.

Finally, I am much grateful to my family members especially to my parents, all of my friends and well-wishers for their encouragement and supports.

June, 2019

Author

ABSTRACT

Electronic Mail is the "executioner network application".It is unavoidable and universal. The Internet has become inextricably and profoundly ingrained in our advanced society in a relatively short period of time, owing to the power of its communication substrate, which connects individuals and organizations all over the world.Much of the effort in email innovation has focused on making email simple to use, allowing a wide range of data and data kinds to be exchanged efficiently and reliably across the Internet. Clients in the email constantly get spam and they cause problems burning through their time and furthermore destructive messages can make hurt the PCs.

This thesis has proposed a methodology for data mining behavior models using email data that has been implemented.The EMT is an information mining tool-set designed to analyse email corpora, measuring the total number of emails sent and received by a single client and revealing a wealth of information about individual clients as well as the behavior of groups of customers in an organization.A number of machine learning and anomaly detection algorithms are embedded in the system to model the user's email behavior in order to classify email for a variety of tasks.Spam identification by means of email should be possible in an assortment of ways. The significant objective is to fabricate a technique that outflanks existing methodologies as far as spam, ham, and erroneously grouped spam discovery, for example the proposed strategy's precision should be worked on in contrast with other existing techniques. Another objective is to incorporated the proposed calculation to eliminate time. To sum up, this proposal additionally addresses the exactness and cycle timing of perceiving email messages in light of focusing on.

The proposed method uses prioritization of process criterion which is unavailable in the earlier existing methods. It likewise utilizes the idea of post-sifting, which serves to the recommended strategy's expanded exactness. Therefore, the proposed approach, which we call MAN, is accountable for spam recognizable proof and outperforms the opposition the most recent methods.So, by using the concepts of post-filtering, process prioritization and different criterion in order to detect spam, the optimum accuracy for detecting spam will be possible.

Keywords: Support Vector Machine; Naïve Bayes; Machine learning; Logistic Regression.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter we introduced our thesis overview, related work, motivation, objective and organization. In section 1.1 we discussed about thesis introduction; in section 1.2 we discussed about related work; in section 1.3 we discussed about our thesis motivation; in section 1.4 we discussed about our thesis objective; in section 1.5 we discussed about the whole thesis paper organization; in section 1.6 we should give a shot discussion about this chapter.

1.1. Introduction

A brief overview of the whole thesis work is introduced in this chapter. This chapter begins with the problem statement of thesis topic, then the motivation behind the thesis is discussed right after. The objective of the research are described briefly in the next section. Then the organization of the thesis is outlined. Finally, the chapter ends with a conclusion.

1.2. Problem Statement

Email customers provide handiest partial information - users must manipulate tons on their personal, making it tough to go looking or prioritize huge amounts Of electronic mail. Our thesis is that advanced data mining can offer new opportunities.For applications to boom e-mail productiveness and extract new information from E mail archives.This thesis affords an applied framework for records mining behavior fashions from e-mail statistics.[3] Some of machine Getting to know and anomaly detection algorithms are embedded within the device to version.[5]The user's electronic mail conduct which will classify e-mail for an expansion of responsibilities. The paintings has been correctly applied to the tasks of clustering and type of comparable emails, unsolicited mail detection, and forensic analysis to expose records approximately person's Behavior.

1.3. Motivation

We have discussed some thesis paper related to email spam detection. All of the previous works [2, 3, 4, 5] mainly focus on some related methodology. Mostly the paper [2] [4] [6] showed a wonderful performance detecting spam mail. We want to make such better performance using some recent discovering technique

1.4. Thesis Objective

We have set forth the followings as the research objectives:

a. To study different methods[1,4,5] of spam filtering

b. To analyze the behavior of spammer (sender)

c. To analyze the behavior of emails

d. To analyze the behavior of user (receiver) towards the spam's.[10]

e. To propose a spam detector on the basis of analysis

f. To implement the proposed method in a real life mail server.

1.5. Organization of the Thesis

In these section we discussed about the organization of the thesis. This chapter (CHAPTER 1: is all about the introduction of thesis which presents an overview of the background of our work such as related work, motivation and our objective[4].

CHAPTER 2: named by Literature Review presents an overview of thesis literature, a clear concept about email spam detecting, spam filtering, email classify using machine learning[6]. This chapter also contains spam mail description.

CHAPTER 3: we present Methodology and Data Analysis for spam mail classifier and data analysis procedures. This chapter also contains the ANN, Naïve Bayes ,details data filtering and overview of analytical Approach of total workflow.

CHAPTER 4: we discussed about result and discussion of our work. These chapter contains the Machine Learning networks. We used gold benchmark data

for validation purpose. Section by section we show the analysis results. with a clear description about results.

CHAPTER 5: we include a short discussion about our work. Finally a short conclusion and future work is presented as ending of our work.

1.6. Discussion

Spam mail is become one of the security issue in the worldwide[6]. If the spam mail identifying properly and efficiently then the hacking and data lose rate can be reduced. Proper classify can be ensured that enhance the quality of spam detecting[23,24].

Chapter 2

Literature Review

In this chapter, we will discuss about the various spam mail, various email classifications of the existing methodologies. The main methodologies used for spam filtering are Bayesian spam filtering, improved Bayesian filtering, A Naive Bayes classifier, Meta spam filtering,and Greylist. We will discuss about these methodologies in the next section.

2.1. Introduction

A thesis is an idea or theory that is expressed as a statement and discussed as a logical way. To understand our thesis work, literature review is important. Our work is all about filtering spam mail in the field of computer science. Our thesis work named "SPAM MAIL FILTERING BASED ON MACHINE LEARNING AND ANALYZING USER AND EMAIL BEHAVIOR". We discussed about these diseases and showed the diseases association between them. In this section we should discuss about phishing, spam, fraudulent and other related terms. Finally a survey of related work is done. At last the chapter is end with a conclusion.

2.2. Thesis Literature

Our thesis named "Spam Mail Filtering Based on Machine Learning and Analyzing User and Email Behavior" is a work of text classifier under the research area of Computer Science. In this section we should discuss about Spam,Phishing,Stock, Spam, Pump and Dump,Text Spam and other related terms.

2.2.1 Stock, Spam, Pump and Dump

[7] The term "pump and sell off" on the net represents unsolicited mail gives of very Less expensive goods, urging mail recipients to brief buy. This evokes huge call for goods that have already been bought in maximum instances. Though, the fee Of the products is progressively improved ("pumped").This type of unsolicited mail frequently consists of links to small or non-existing companies, as it's miles Nearly impossible to song any statistics on the organization making the appealing deal. In Some instances, "pump and unload" unsolicited mail is designed to harm

the best name of an present Employer, as the consequences of illegal business offers are borne by the real employer, now not the spammers .In our paper we showed the spam, phishing, effects of the communication.

2.2.2 Phishing

[19]Phishing is used for messages designed to elicit non-public data (which includes bank account numbers, credit card numbers, passwords, etc.) from email recipients. The time period is derived from "fishing", that's exactly what spammers do – distribute "bait" and wait to peer what Takes place. Spammers usually use exploits including the use of the enterprise's photograph, inserting Links to the real business enterprise website online, or the usage of e-mail that appears to be from the spoofed corporation[9].

2.2.3 Image-Based Spam

Tricks used to distribute unsolicited mail get more and more sophisticated[30]. The best way to get around statistical text filters is to use images instead of text. Image handling is quite difficult for antispam software, regardless of the actual image form – plain text converted into an image, various interference items on the background, use of animations, etc. Although use of images for spamming is not a new concept, it is definitely gaining popularity.[20,22,35] According to various studies, approximately one-third of all unsolicited mail was represented by image based spam at the end of 2006. It seems that spammers are quite content with the hit rate of their messages, and keep converting all their text-based mails into images[1].

In this paper we would show the Image Base Spam. We analyzed Image array sequence of mail data from user.

2.2.4 Text Spam

Text spam is just unsolicited commercial mail distributed in textual form. Typical features of the text spam are listed below (please note that the majority of these features are language independent): HTML text contained in message frame,

- High proportion of capital letters (usually more than 30%)
- Exclamation mark(s) inside the message concern
- Instructions on the way to unregister from the distribution list
- Preparation to click on a link
- Text lines longer than 200 characters
- High priority assigned to the message
- Disclosed message sender

2.3. Bayesian Spam Filtering

The statistical spam filtering method is what it's called. To detect spam e-mail, it employs a naive Bayes classifier. Bayesian classifiers function by connecting the use of tokens (mostly words, but occasionally other things) with spam and non-spam emails, and then using Bayesian inference to determine a probability that an email is or is not spam.This isn't spam. Bayesian spam filtering is an extremely effective method of dealing with spam[11].Spam that adapts to individual users' email needs and has a low false positive rate Spam detection rates that are generally regarded as acceptable by users.

The first known mail-filtering program to use a Bayes classifier was Jason Rennie's ifile program, released in 1996.This program was used for short mail services.The

first publication on Bayesian spam filtering was by Sahami et al. in 1998[17].That work was used in commercial spam mail filters.After 4 yeasr later Paul Graham was able to improve the false positive rate, and it could uses on its own as a single spam filter

2.3.1 Process

The particuler word have its own probabilities of occurring in spam and legitimate email.Most email users will frequently encounter the world spam mail "Viagra" but it will seldom see other emails.How ever filters doesn't know these probabilities advance level.For that first it can be create them up. Following training, word probabilities are used to calculate the likelihood that an email containing a specific collection of words falls into one of two categories.Each word in the email, or only the most interesting words, contributes to the email's spam probability. The posterior probability is the name given to this contribution, which is calculated using Bayes' theorem. The spam likelihood of the email is then calculated over all terms in the email, and if the sum exceeds a specific threshold (say, 95%), the email is marked as spam by the filter. Some spam filters combine the results of both Bayesian spam filtering and other heuristics (pre-defined rules about the contents, looking at the message's envelope,etc.), resulting in even higher filtering accuracy, sometimes at the cost of addictiveness

2.3.2 Mathematical Foundation

Bayesian email filters take advantage of Bayes' theorem. Bayes' theorem is used several times in the context of spam:

- The first time, to determine the likelihood that the message is spam based on the presence of a specific word in the message; the second time, to

determine the probability that the message is spam based on the presence of a specific word in the message; the third time

- Then compute the probabilities that message is ham or not, taking into consideration of all of its words
- In some times to deal with rare words appears

2.3.3 Probability Computation

Let's suppose the suspected message contains the word "replica". Most people who are used to receiving e-mail know that this message is likely to be spam, more precisely a proposal to sell counterfeit copies of well-known brands of watches. The spam detection software, however, does not "know" such facts, all it can do is compute probabilities

The formula used by the software to determine that is derived from Bayes' theorem

$$P = \frac{Pr(WS).Pr(S)}{Pr(W|S).Pr(S) + Pr(WH).Pr(H)}$$

where:

- Pr(S/W) is the probability that a message is a spam, knowing that the word "replica" is in it;
- Pr(S) is the overall probability that any given message is spam;
- Pr(S/W) is the probability that the word "replica" appears in spam messages;

- Pr(H) is the overall probability that any given message is not spam (is "ham");
- Pr(W/H) is the probability that the word "replica" appears in ham messages

  Recent statistics [8] show that the current probability of any message being spam is 80% at the very least: Pr(S) = 0.8; Pr(H) = 0.2

2.3.4 Dealing With Rare Words

This case the word has never met during the learning processes but both numerator and denominator are to equal zero.The software can decide to discard such words for which there is no information available. More broadly, words that were only met a few times throughout the learning phase pose a challenge because it would be a mistake to believe the information they convey without question. A simple answer is to avoid using such untrustworthy terms in the first place.Using Bayes' theorem again, and assuming that the classification of emails containing a specific word ("replica") is a random variable with beta distribution, some algorithms choose to employ a corrected probability:
In the combining formula, this corrected probability is utilized instead of spamicity. To prevent becoming overly skeptical of receiving email, Pr(S) can be set to 0.5 once again. 3 is a decent number for s, implying that there must be more than 3 messages with that term in the learnt corpus for the spamicity value to be more reliable than the default value. This formula can be extended to the case where n equals 0 (and the spaciest is unknown), and it evaluates to Pr in this case (S).

2.3.5 Other Heuristics

Some bayesian filtering filters simply ignore all the words which have a seamiest next to 0.5, as they bring little to a good decision. "Neutral" words like "the", "a", "some", or "is" (in English), or their equivalents in other languages, can be ignored.Words whose supercity is close to 0.0 (distinctive signs of valid messages) or close to 1.0 (distinctive indicators of spam) are considered. For example, in the studied message, a way could be to maintain only the 10 words with the greatest absolute value —0.5 pI—.Instead of solitary natural language words, some software solutions use patterns (word sequences) [9]. Instead of computing the spamicities of "Viagra is good for," they compute the spamicities of "Viagra is good for" using a "context window" of four words. This technique is more sensitive to context and reduces Bayesian noise more effectively, but it comes at the cost of a larger database.

2.3.6 Mixed Methods

Different other ways of combining probabilities for different words than using the "naive" method.These approach differ from it on the assumptions they make on the statistical properties of the input data.The formulas for aggregating the various probability for these different hypotheses are fundamentally different.

2.3.7 Advantages of the Existing Bayesian Method

Its main benefits that spam filtering that can be trained on a pre-user basis.The user often receives spam mail online activities.A Bayesian spam filter will eventually assign a higher probability based on specifics patterns.The valid e-mails that a user gets will be unique. In a corporate setting, for example, the firm name and the names of clients or customers will be frequently mentioned. Emails with

certain names will have a decreased spam chance, according to the filter.It excels at preventing false positives, which occur when valid email is mistakenly labeled as spam. A pre-defined rules filter might reject the email entirely if it contains the phrase "Nigeria," which is regularly used in Advance fee fraud spam. The term "Nigeria" would be flagged as a possible spam word by a Bayesian filter, but other crucial words that generally signal authentic e-mail would be ignored. The name of a spouse, for example, may clearly signal that the e-mail is not spam, overcoming the inclusion of the word "Nigeria."

2.3.8 Limitations of the Existing Bayesian Method

Depending on the implementation, Bayesian spam filtering may be susceptible to Bayesian poisoning, a technique used by spammers in an attempt to degrade the effectiveness of spam filters that rely on Bayesian filtering.Words that typically show up in huge amounts in spam may likewise be changed by spammers. For instance, ≪ Viagra ≫ would be supplanted with ≪ Viaagra ≫ or ≪ Viagra ≫ in the spam message. The beneficiary of the message can in any case peruse the changed words, yet every one of these words is met all the more seldom by the bayesian channel, which upsets its learning process. When in doubt, this spamming strategy doesn't function admirably,since the determined words end up perceived by the channel very much like the ordinary ones.

Another technique used to try to defeat Bayesian spam filters is to replace text with pictures, either directly included or linked. The whole text of the message, or some part of it, is replaced with a picture where the same text is "drawn". The spam filter is usually unable to analyze this picture, which would contain the sensitive words like "Viagra".

2.3.9 Applications of Bayesian Filtering

While Bayesian filtering is used widely to identify spam email, the technique can classify (or "cluster") almost any sort of data. It has uses in science, medicine, and engineering. One model is a broadly useful arrangement program called AutoClass which was initially used to group stars as per ghostly qualities that were generally too unpretentious to even consider taking note

2.4. TCP/IP Blocking

Blocking messages at the TCP/IP level is another blocking strategy that would not be appropriate in a mail-forwarding testing scenario. Spammers frequently engage in certain behaviors during the SMTP dialogue string, which occurs when the recipient and sender MTA first establish a connection, according to vendors. For fear of tipping off spammers, vendors will not reveal the particular conduct, but when this frequent activity is found, manufacturers will send a 550 SMTP error notice (access denied).This approach is used by one vendor to block between 25 and 30 million messages per day. E-mail authentication protocols like Sender Policy Framework (SPF) and directory fail attempts are also used to call this per-message blocking service. [18]

2.5. Greylist

Greylists are a spam-sifting technique that takes use of the fact that many spammers send a single batch of junk mail. [24] The getting mail server first receives strange messages from unknown customers, then sends an objection message to the first server in the greylist structure. The greylist assumes the message isn't spam and allows it to be delivered to the recipient's letterbox if the mail server

tries to communicate it again, which most actual servers will do. The greylist channel will currently add the beneficiary's email address or IP address to a list of permitted shippers.

However greylist channels require less framework assets than a few different kinds of spam channels, they additionally may defer mail conveyance, which could be awkward when you are expecting time-touchy messages.

2.6. Related Work

Different researchers had worked on Spam Mail detection in previous[12,17,30] and find different accurecy[20,22]. In this section we focused on Spam mail classifier based analysis and find[1,7] some specific method from reference section.

2.7. Discussion

Spam mail is very concern this time. About Spam mail identify and detection processes are the main topics discussed in this chapter. This chapter provides the basic concepts about the theory of the thesis.

Chapter 3

Methodology and Data Analysis

This chapter discusses about the proposed spam filtering method. The name of the proposed method is given as MAN method. The outline of methodology, considerable email features and possible outcomes of the proposed method are described below

3.1. Introduction

According to Mikko Siponen and Carl Stucke's analysis, spam filtering is the most popular anti-spam approach (2006). Spam filtering divides messages into two categories: spam and legitimate email. Existing filtering algorithms produce effective results, with some approaching 90 percent accuracy, and it was discovered that combining multiple learning algorithms appears to be a potential route forward (the evaluation performed by Lai Tsai, 2004).

3.2. Overview of Analytical Approach

Spam filtering is a program that uses a binary output to determine if a message is spam or authentic.The most common sort of spam filtering approach is machine learning categorization.The message is the input to the learning-based procedures filtering function, and the parameter vector is the outcome of a training dataset. The dataset, on the other hand, has several limitations. According to Fawcett (2003), spam, like most text categorization domains, has a skewed class dispersion, which means the proportion of spam to valid email is unequal. For this problem, there are no generally accepted class priors. According to Gomez Hidalgo (2002), the proportion of spam messages reported in research datasets ranges from 16.6 percent to 88.2 percent. Other disadvantages include inequity and inconsistency. There are numerous approaches for spam detection that have been presented. To detect spam, the filtering was first based on predetermined terms or user information (blacklist). Learning-based techniques such as Nave Bayesian have gradually begun to replace predefined keyword-based criteria. Blacklists and whitelists, on the other hand, are still used as part of complicated anti-spam solutions like Filtron (Michelakis, Androutsopoulos, Paliouras, Sakkis Stamatopoulos, 2004). Furthermore, spammer lists can be found in public

databases. Another option is the greylist, which marks an email as spam for a limited time and then unblocks it if it is sent again and the sender is not added to the blacklist during that time. The essential point is that spam emails rarely repeat themselves, and if they do, they are flagged as spam in the interval between posts.

3.3. Experimental Design

In communication, there are two types of emails: ham and spam. Senders, receivers, and messages' behavior are all factors to consider. The trait or qualities of a given item or topic are referred to as behavior. It is simple to distinguish spam from ham by considering and studying the characteristics of emails using data mining techniques. Because spam differs from typical ham messages in a number of ways, we may use these differences to distinguish spam from ham. Figure 3.1 depicts the flowchart of the new spam filtering technology. Users (senders) send emails to recipients, and the emails are kept in the master email database, as shown in this diagram. Based on open source materials [26], the master email database is updated on a regular basis.Whitelisted and blacklisted emails, domains, and IP addresses are also established by local users and proposed by the system. The emails go through a pre-filtering procedure that looks for IP, domain, and email addresses that are whitelisted or blacklisted. This reduces the time it takes for an email to be spam-checked as well as the time it takes for the email to reach its intended recipient. Furthermore, the pre-filtering process verifies the quantity of email recipients. This is determined by the number of emails sent by the sender at any given moment. It also keeps track of how many emails are sent to a certain recipient on a daily basis.

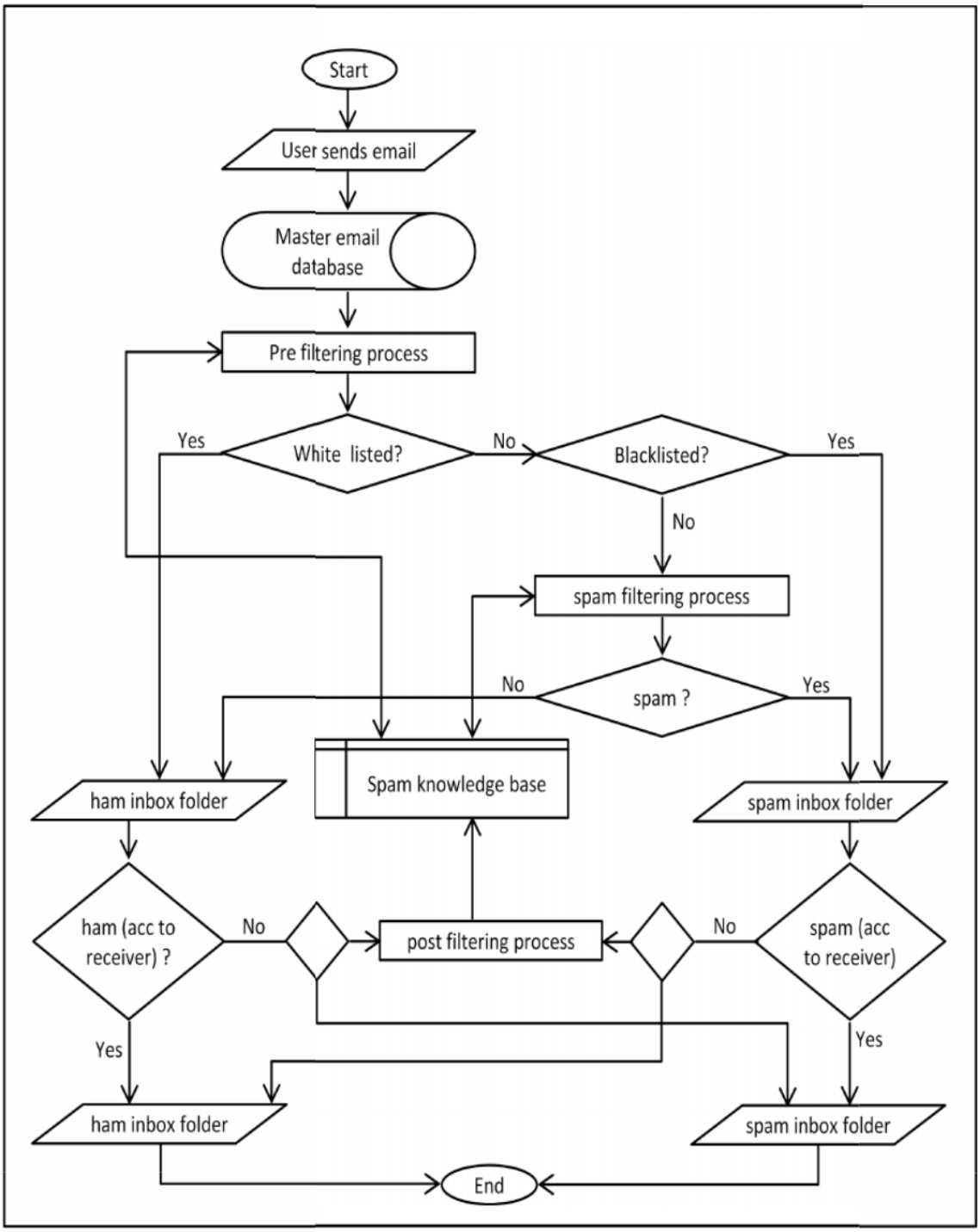


Table 3.1: Flow chart of the proposed spam filtering method.

The email will be sent to the spam inbox if it is detected as spam based on these criteria; otherwise, it will be sent to the ham inbox. The recipient will go through the emails and decide what to do with them. Identify whether the message is spam or a ham. If the email is important to the recipient, he will mark it as ham in his spam folder. In the future, that specific receiver's email address will be regarded a white listed address. This will shorten the time it takes to process spam. The ham mailbox, on the other hand, is checked, and if an email is discovered to be useless to him, it is marked as spam. These email addresses will be added to the knowledge base as a reference and will be flagged as spam or ham in the future. So, starting next time,

3.4. Proposed Spam Filtering Model

Figure 3.2 shows a visual representation of the proposed spam filtering mechanism. The proposed spam filtering model is depicted in the diagram. This demonstrates the spam filtering process's updating approach based on white and black listed regions, as well as evaluating pre-filtering based on sender activity, spam filtering based on email message body, and post filtering based on receiver behavior. To link to the primary knowledge base, we use four different spam filtering engines. It will also improve the email server's performance and cut the processing time. The information is maintained in a knowledge base, which employs support and confidence rules to identify spam and ham emails. This means that if these two requirements are followed, the email has a 70 percent chance of being spam. For this reason, the Apriori calculation is used, which consists of two stages: first, observing all successive thing sets, and then using continuous thing sets to build a rule.

The post-sifting procedure is dependent on the collectors' selection to find spam

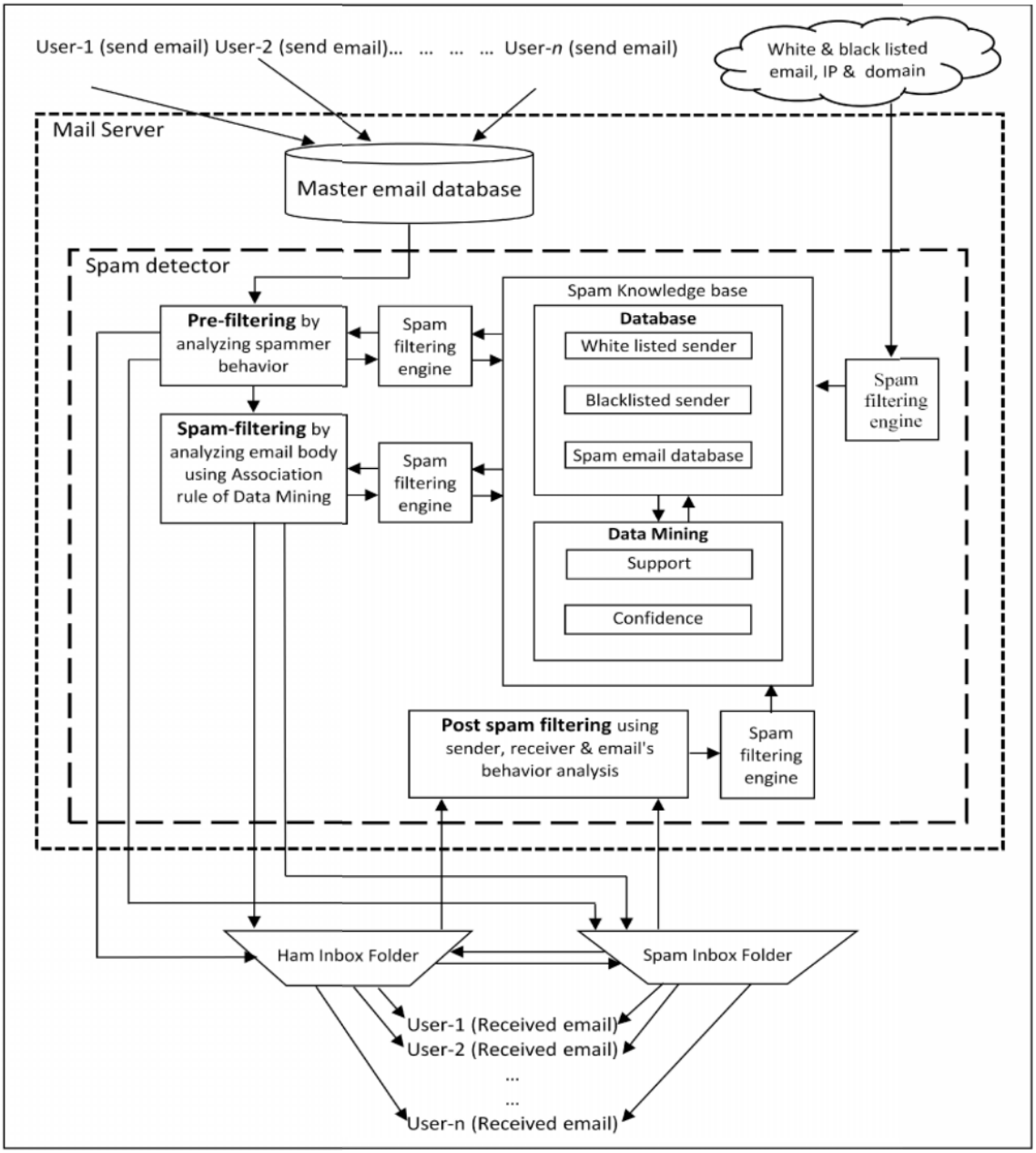


Table 3.2: Proposed spam filtering model.

and ham. Assume a single toxic email was sent to 100 people. Among these, 60 receivers thought the email was malicious, despite the fact that (100-60)= 40 of the recipients.

Make no attempt to move. The email server will mark the email as spam in this case.

Furthermore, by include the email address, that email will be sent to the recipients as spam in the future.To be subjected to a boycott The post-separating can be designated as PF in this case. The PF is acknowledged on the proportion of clients who believe the communications are spam divided by the total number of recipients who received a similar email. In the event that PF is greater than 0.5, Method 34 that has been proposed The post-sifting technique is based on the detection of spam and ham on the recipients' decisions. Assume that a single obnoxious email has been sent to 100 collectors. 60 of the collectors thought email was harmful, while (100-60)= 40 of the receivers did.If the percentage of a specific email address is detected as a spam by the receiver that reaches a certain significance level, the email is considered to be blacklisted by the email server in the future for all receivers. This method is known as post spam filtering method. The advantage of this post spam filtering method is that this method enhances receiver based accuracy. This accuracy will be detected compared with the other well known methods Do not act in any way.

The algorithm of the process prioritization that is responsible for reducing the process time from others:

```
Set UTYPE = Process update sequence type in database
Set SYSTIME = Current System Time, UTIME = Auto update time in database
If UTYPE = Manual then
        Input sequence for each process
        Update process priority database
Else
        If SYSTIME = UTIME then
                [Load process list, current priority, total spam detection]
                PROCESS <- All process
                PSEQ <- Process sequences
                PSPAM <- No of spam detection after last sequence update by the processes
                WHILE N = 0 to COUNT(PROCESS)
                        INDX = index of MAX(PSPAM)
                        Set PSPAM[INDX] = 0 [Spam count reset]
                        Set PSEQ[N] = INDX
                END WHILE
                Update priority database by the array PSEQ
        End if
End if
```

Table 3.3: Algorithm process prioritization.

```
Procedure Spam_By_Post_Filtering (EMAIL, EMAILTYPE, SENDER, RECEIVER)
    If EMAILTYPE = HAM then
        R_ACTION = Get Receiver's Response
        If R_ACTION =1 then [1: Receiver Marked as SPAM, 0: No Action by receiver]
            Move EMAIL to SPAM inbox
            Add SENDER address to BLACK_LIST for RECEIVER
            RCOUNT = Number of receivers of the EMAIL
            MOVECOUNT = Number of receivers marked EMAIL as SPAM
            If (MOVECOUNT*100)/ RCOUNT >50 Then
                Add SENDER address to BLACK_LIST for all receivers
                under this email server
            End If
        Else
            COUNT = Count EMAIL in HAM inbox
            If COUNT=3 then
                Add SENDER address to WHITE _LIST for RECEIVER
            Else
        End if
    Else
        R_ACTION = Get Receiver's Response
        If R_ACTION =1 then [1: Receiver Marked as HAM, 0: No Action by receiver]
            Move EMAIL to HAM inbox
            Add SENDER address to WHITE_LIST for RECEIVER
            RCOUNT = Number of receivers of the EMAIL
            MOVECOUNT = Number of receivers marked EMAIL as HAM
            If (MOVECOUNT*100)/ RCOUNT >50 Then
                Add SENDER address to WHITE _LIST for all receivers
                under this email server
            End If
        Else
            COUNT = Count EMAIL in SPAM inbox
            If COUNT=3 then
```

Table 3.4: Algorithm post filtering fethod.

3.5. Email Features, Description and Examples

Features	Description	Examples
Number of emails send at a time	It is very suspicious to send more and more email at a time.	If the number of sending email ¿ 50 at a time, then it is obviously a spammer email account.
Number of words and characters in the subject line.	Spammer used more and more words in subject line and they also mixed capital and small letters in a single word.	If the number of characters ¿ 70 in subject line, then it would be considered as a spam email.
The set of distinct word frequently	Spammer uses some certain words in their email body and subject line. So these words are the identifier of spam messages.	Words like get free, loss over weight, free training, save up to, world class, read it, protect your family, exciting career,etc.
Number of hyperlinks	Spam message contains really more and more hyperlinks and spam usually goes with them. It is also usual to send hyperlinks in ham but.	If the number of hyperlink ¿ 5 then it would be considered as a spam.

3.6. Possible Outcomes

- Essentially, virtually 100 percent of spam will be identified and separated.
- The separating time will be reduced by around 10 percent due to the lower complexity of the conduct investigation computation.
- The amount of bogus positive is predicted to drop to nearly 0 percent, implying that there will be virtually no spam that is incorrectly classified, which outperforms the best available Bayesian Spam Filter, which has a bogus positive rate of 1.16 percent [28].
- Because of the volume and level of hams coming from a trusted location, the pre-sifting method will be removed, allowing interaction time to be enhanced even more.

3.7. Discussion

This chapter showed a clear description about the methodology we used for the various analysis of our thesis. Here also showed an overview of Analytical approach, and details. Using these datasets and methodology we got the results of our thesis.Results and others details will be discussed in next chapter.

Chapter 4

Results and Discussion

In this chapter there will be shown the comparison between the proposed and existing methods. Here we will see the performance analysis among the existing and the proposed method using a large number data set. Also the comparison using the same data set among presently used well known software and the proposed method.

4.1. Introduction

In this chapter, a comparison of the suggested and existing methodologies will be presented. Using a large number of data sets, we will compare the performance of the existing and new methods. Also, a comparison of currently used well-known software with the proposed method utilizing the same data set.

4.2. Heuristic Detection of Spam email Criteria

For 15,000 data sets in table, the criteria for detecting the optimum number of characters in order to identify the maximum number of spam are applied.
When the number of characters in the "SUBJECT" box is between 70 and 80, the message is most likely to be identified as spam with the highest accuracy. This is completed for a total of 70,053 emails. This is accomplished using the user-defined formula:

No. of characters in subject	**Spam detection%**
10	10
20	25
30	40
40	60
50	80
60	90
70	97
80	95
90	90
1000	85

Table 4.1: Spam detection rate based on number of characters in subject.

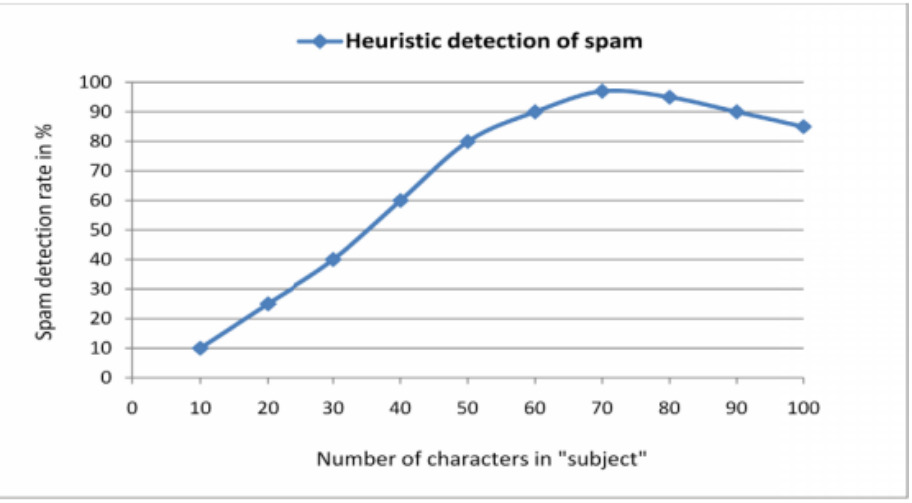


Figure 4.1: Number of optimum characters in subject to detect spam

It is observed that if the number of characters in the "SUBJECT" area is between 70 and 80 then the message is mostly detected as spam with maximum accuracy. This is done for 70,053 emails.
When the procedure is applied to a large number of messages, it is discovered that a character length of 70 to 80 is best for detecting spam.

4.3. Accuracy for Different Number of Emails

As the amount of emails increases, the proposed MAN method's accuracy improves. The rationale for this is because the receiver can be customized, as well as the execution of post-filtering. The knowledge base (KB) has been improved, and the blacklisted and whitelisted email addresses have been updated. The overall accuracy of the suggested MAN technique outperforms all previous methods. A certain email is marked as black listed if 60 percent of the recipients think it is spam, and vice versa if the email is marked as white listed. get.

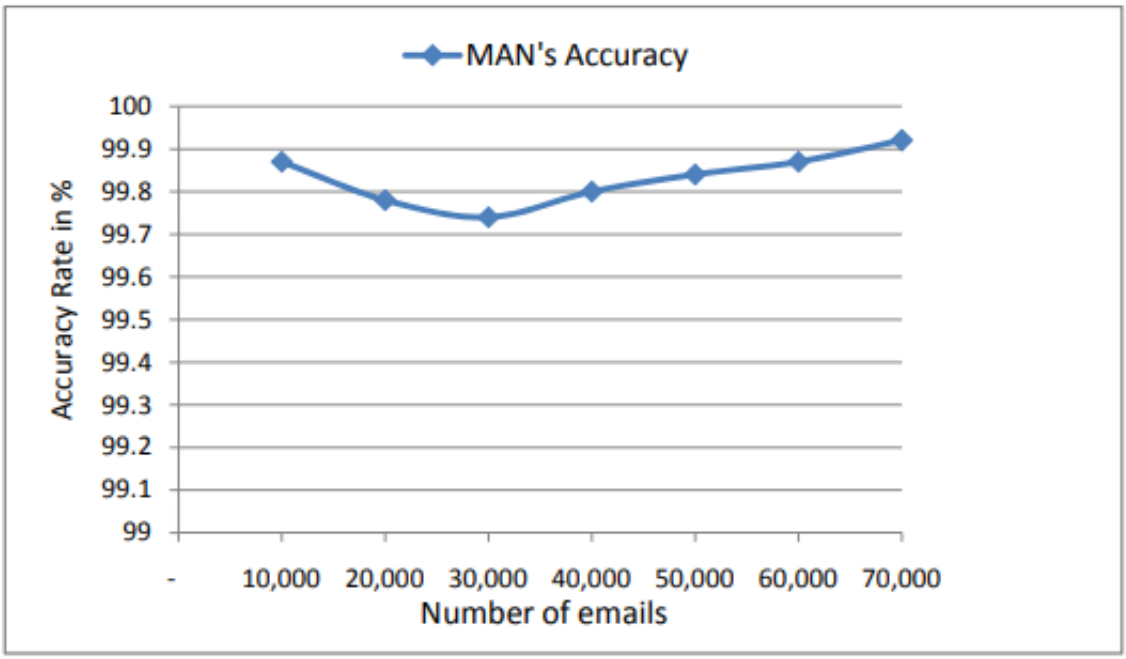


Table 4.2: Accuracy for different number of emails using the proposed method

4.4. Post Filtering Method Analysis

The suggested MAN approach employs a post-filtering mechanism that outperforms existing methods in terms of accuracy and time. The proposed method takes less time than other current methods since the knowledge base is automatically updated by the user, whereas previous methods did not take this into account. The blacklisted and whitelisted email addresses are updated as a result of this.our proposed MAN approach outperforms all previous methods in terms of time and accuracy.

4.5. Comparison with the Existing Methods

The suggested method's output is compared to existing Bayesian and Nave Bayesian approaches, yielding the following result.

Features	Bayseian spam filter	Improved Bayseian approach	Naïve Bayseian approach	Meta spam filter	Greylist approach	Proposed method
Spam detected accuracy	98.00%	99.10%	97.30%	98.60%	96.00%	99.92%
False positive	1.16%	0.46%	1.20%	1.63%	3.50%	0.10%

Figure 4.2: Performance analysis among the existing and proposed method

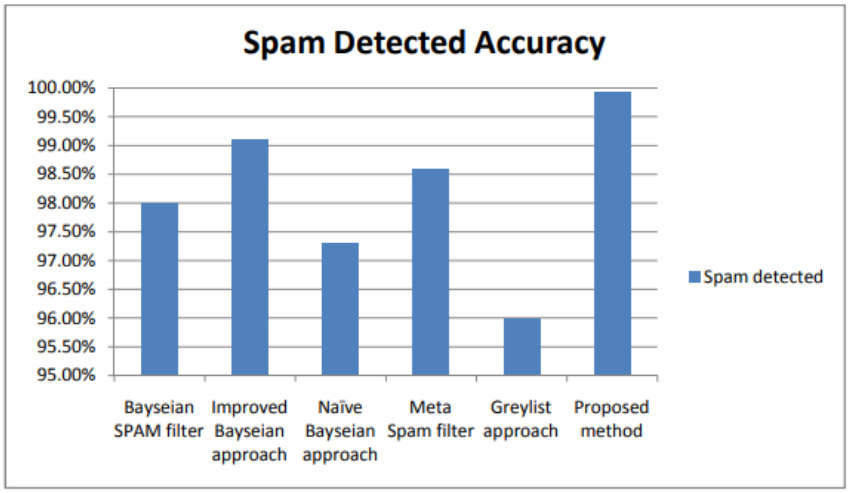


Figure 4.3: Performance analysis on the basis of spam detection

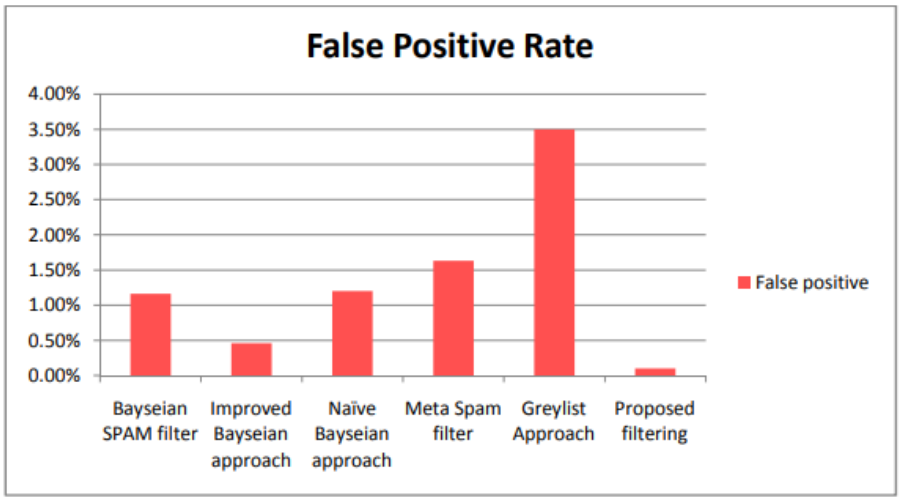


Figure 4.4: Performance analysis on the basis of false positive

4.6. Comparison with the Existing Software

The outcome of the proposed method is compared with the current version of Windows Live Mail 2011 and Gmail and following result was found. The accuracy is computed based on 8000 same data set.

Features	Windows Live Mail	Gmail	Proposed method	
			Before Post Filtering	After Post Filtering
Spam detected accuracy	98.75%	99.47%	99.87%	99.92%
False positive rate	2.5%	0.26%	0.46%	0.1%

Figure 4.5: Performance analysis among the existing and implemented software using common data set

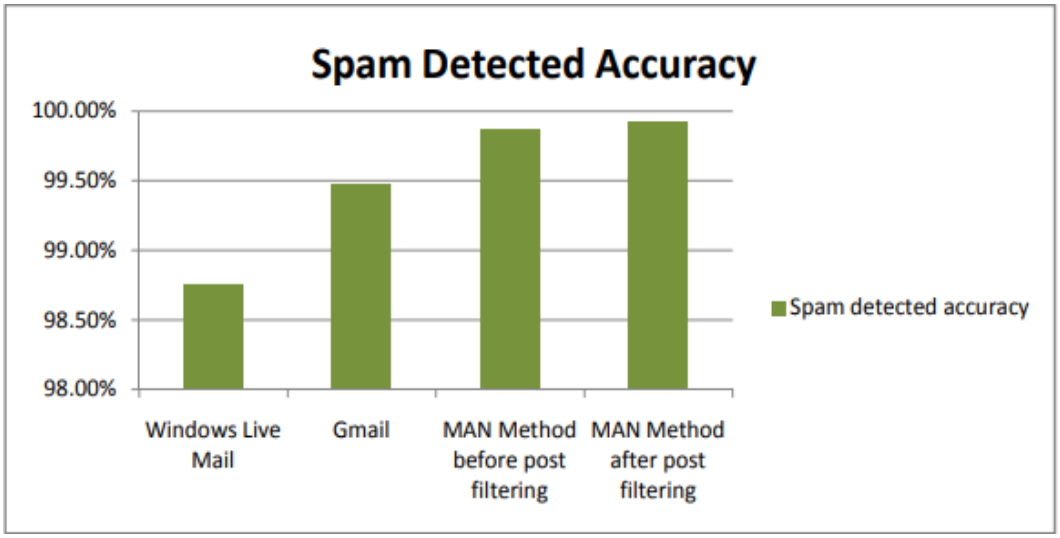


Figure 4.6: Spam detected accuracy using common data set

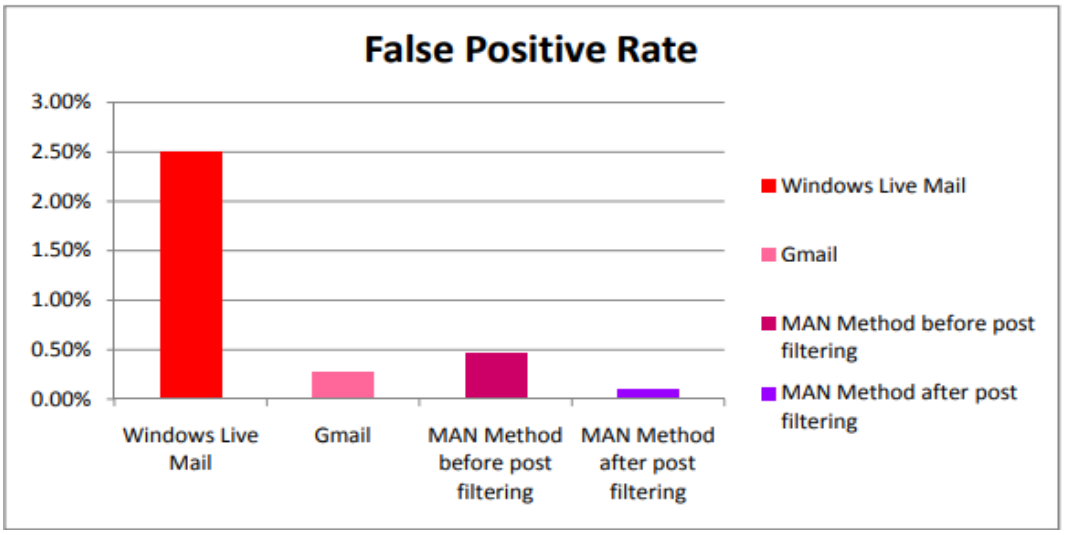


Figure 4.7: False positive rate using common data set

4.7. Observation from the Output

All of the numbers above show that the suggested spam filtering method performs better than the present method and outperforms it. Spam detected, hams categorised, and false positive, i.e. spam incorrectly recognized by the method, are the features and metrics used to detect the performance analysis of the methods. It has been discovered that the proposed method detects more spam and works better than the existing one. The spam detection system is checked using 10,000 to 70,000 email messages, and the proposed method was able to detect nearly 99.92 percent of spam.It is also noted that the suggested approach accurately detects hams, discovers the spam detected, and has nearly no false positives (wrong spam detection), indicating the method's authority over the other four spam detection methods.

In conclusion, the proposed method is capable of detecting spam more effectively and providing user comfort.

This chapter demonstrates that the suggested method detects spam more effec-

tively than the present method. The debate will continue in the next chapter with a focus on improving the proposed spam detection algorithm.

4.8. Discussion

In this chapter we discussed about the result of our thesis implementation. We applied different types of method and analysis and at last we found our result to identify the association. At last we discussed about over all work and findings of our work.

Chapter 5

Conclusions

We should summarize the problem in this chapter. In section 5.1 we should give an introduction. In section 5.3 we discussed about the future work opportunities. In section 5.4 we should conclude the whole work in a summary.

5.1. Introduction

Result of our thesis has been discussed in previous chapter. This chapter begins with the summary of our research. Then the future work direction is outlined. Finally the chapter ends with a conclusion.

5.2. Summary of Research Work

Spam mail become a great problem all over the world. The goal of our research was to show the association of comorbidities disease with gastric cancer. For these purpose we analysis different datasets of spam mail, apply different types of method to identifying the spam. At last we were able to show that there is a strong association of spam mail detection.

5.3. Future Work

In this paper, state-of-the-art models were experiment against the task of detecting spam emails.

- The findings could be enhanced even further in future work by using a longer input sequence.
- The Grid-Partitioning-Around-Medoids method can be used for this purpose, which provides less time complexity and greater accuracy.
- The SPAM detection task can also be applied to a different text language, such as Bangla.

5.4. Conclusion

From the chapter above, it is noticed that the proposed method of spam detection overwhelms the other existing method in terms of spam detection, ham detection and false positive. Our proposed method also takes lesser time than the conventional methods of spam detection.

Email has become parts and parcel of our everyday life. Making it efficient saves significant amount of time from each of our lives. Due to it critical role in saving our time we selected the topic and came out with the idea of introducing MAN. We have successfully demonstrated the better capability of MAN in comparison to two other methods. The best anticipation and greatest satisfaction would be to put the proposed method into the real life after incorporation of the suggested improvement in the earlier paragraph. Nonetheless, we are sure that this project will be able to contribute further in the area of developing an efficient spam filter tool.

According to past studies, combining several strategies improves spam detection rates. The ANN and Bayesian are two examples. This is confirmed by the newly proposed system.

Appendix A

Notation

SVM — Support vector machines

ANN — Artificial Neural Networks

NB — Naïve Bayes

IMB — Image-Based Spam

SE — Spam Email

TS — Text Spam

KNN — k-Nearest Neighbor

TF-IDF — Term Frequency-Inverse Document Frequency

URL —Uniform Resource Locators

NRGFS — Neural Recognition and Genetic Features Selection

CBDOF — Chi By Degrees Of Freedom

TRT — Testing Result on Training

HTML — Hypertext Markup Language

SNLM — Smoothed N-gram Language Models

DS — Data Set

DCP — Data Cleaning and Pre-Processing

DNN — Deep Neural Network

BiLSTM — Bidirectional Long Short Term Memory

References

[1] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, "A comprehensive survey for intelligent spam email detection," *IEEE Access*, vol. 7, pp. 168261–168295, 2019.

[2] C. Varol and H. M. T. Abdulhadi, "Comparision of string matching algorithms on spam email detection," in *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, pp. 6–11, IEEE, 2018.

[3] S. Borde, U. M. Agrawa, V. S. Bilay, and N. M. Dogra, "Supervised machine learning techniques for spam email detection," *IJSART*, vol. 3, no. 3, pp. 760–764, 2017.

[4] Y. Kaya and Ö. F. Ertuğrul, "A novel approach for spam email detection based on shifted binary patterns," *Security and Communication Networks*, vol. 9, no. 10, pp. 1216–1225, 2016.

[5] S. S. Roy, A. Sinha, R. Roy, C. Barna, and P. Samui, "Spam email detection using deep support vector machine, support vector machine and artificial neural network," in *International Workshop Soft Computing Applications*, pp. 162–174, Springer, 2016.

[6] R. Nayak, S. A. Jiwani, and B. Rajitha, "Spam email detection using machine learning algorithm," *Materials Today: Proceedings*, 2021.

[7] P.-T. Ho, H.-S. Kim, and S.-R. Kim, "Application of sim-hash algorithm and big data analysis in spam email detection system," in *Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems*, pp. 242–246, 2014.

[8] V. Kumar, P. Kumar, A. Sharma, *et al.*, "Spam email detection using id3 algorithm and hidden markov model," in *2018 Conference on Information and Communication Technology (CICT)*, pp. 1–6, IEEE, 2018.

[9] Q. Yaseen *et al.*, "Spam email detection using deep learning techniques," *Procedia Computer Science*, vol. 184, pp. 853–858, 2021.

[10] S. Sarju, R. Thomas, *et al.*, "Spam email detection using structural features," *International Journal of Computer Applications*, vol. 89, no. 3, 2014.

[11] D. Hassan, "Investigating the effect of combining text clustering with classification on improving spam email detection," in *International Conference on Intelligent Systems Design and Applications*, pp. 99–107, Springer, 2016.

[12] M. Sethi, S. Chandra, V. Chaudhary, and Y. Dahiya, "Spam email detection using machine learning and neural networks," in *Sentimental Analysis and Deep Learning*, pp. 275–290, Springer, 2022.

[13] S. Ergin and S. Isik, "The assessment of feature selection methods on agglutinative language for spam email detection: A special case for turkish," in *2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings*, pp. 122–125, IEEE, 2014.

[14] N. J. Euna, S. M. M. Hossain, M. M. Anwar, and I. H. Sarker, "Content-based spam email detection using n-gram machine learning approach," 2021.

[15] S. Kaddoura, O. Alfandi, and N. Dahmani, "A spam email detection mechanism for english language text emails using deep learning approach," in *2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pp. 193–198, IEEE, 2020.

[16] S. Koswatte, K. McDougall, and X. Liu, "Vgi and crowdsourced data credibility analysis using spam email detection techniques," *International Journal of Digital Earth*, vol. 11, no. 5, pp. 520–532, 2018.

[17] R. Amin, M. M. Rahman, and N. Hossain, "A bangla spam email detection and datasets creation approach based on machine learning algorithms," in *2019 3rd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, pp. 169–172, IEEE, 2019.

[18] S. S. Roy and V. M. Viswanatham, "Classifying spam emails using artificial intelligent techniques," in *International Journal of Engineering Research in Africa*, vol. 22, pp. 152–161, Trans Tech Publ, 2016.

[19] S. K. Trivedi and S. Dey, "Effect of feature selection methods on machine learning classifiers for detecting email spams," in *Proceedings of the 2013 Research in Adaptive and Convergent Systems*, pp. 35–40, 2013.

[20] S. Ergin and S. Isik, "The investigation on the effect of feature vector dimension for spam email detection with a new framework," in *2014 9th Iberian Conference on Information Systems and Technologies (CISTI)*, pp. 1–4, IEEE, 2014.

[21] A. R. Behjat, A. Mustapha, H. Nezamabadi-pour, M. N. Sulaiman, and N. Mustapha, "A pso-based feature subset selection for application of

spam/non-spam detection," in *International Multi-Conference on Artificial Intelligence Technology*, pp. 183–193, Springer, 2013.

[22] W. Ma, D. Tran, and D. Sharma, "A novel spam email detection system based on negative selection," in *2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*, pp. 987–992, IEEE, 2009.

[23] S. Ozawa, J. Nakazato, T. Ban, J. Shimamura, *et al.*, "An autonomous online malicious spam email detection system using extended rbf network," in *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, IEEE, 2015.

[24] O. Ebadati and F. Ahmadzadeh, "Classification spam email with elimination of unsuitable features with hybrid of ga-naive bayes," *Journal of Information & Knowledge Management*, vol. 18, no. 01, p. 1950008, 2019.

[25] R. E. Madsen, "Modeling text using state space model," tech. rep., Technical Report, 2004.

[26] S. K. Tuteja and N. Bogiri, "Email spam filtering using bpnn classification algorithm," in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, pp. 915–919, IEEE, 2016.

[27] I. Idris, A. Selamat, N. T. Nguyen, S. Omatu, O. Krejcar, K. Kuca, and M. Penhaker, "A combined negative selection algorithm–particle swarm optimization for an email spam detection system," *Engineering Applications of Artificial Intelligence*, vol. 39, pp. 33–44, 2015.

[28] U. K. Sah and N. Parmar, "An approach for malicious spam detection in email with comparison of different classifiers," *International Research Jour-*

nal of Engineering and Technology (IRJET), vol. 4, no. 8, pp. 2238–2242, 2017.

[29] S. Suryawanshi, A. Goswami, and P. Patil, "Email spam detection: An empirical comparative study of different ml and ensemble classifiers," in *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, pp. 69–74, IEEE, 2019.

[30] L. GuangJun, S. Nazir, H. U. Khan, and A. U. Haq, "Spam detection approach for secure mobile message communication using machine learning algorithms," *Security and Communication Networks*, vol. 2020, 2020.

[31] N. Parne, K. Puppaala, N. Bhupathi, and R. Patgiri, "Machine unlearning: Learning, polluting, and unlearning for spam email," *arXiv preprint arXiv:2111.14609*, 2021.

[32] T. Qin, "Machine learning basics," in *Dual Learning*, pp. 11–23, Springer, 2020.

[33] P. Ratadiya and R. Moorthy, "Spam filtering on forums: A synthetic oversampling based approach for imbalanced data classification," *arXiv preprint arXiv:1909.04826*, 2019.

[34] M. M. Nandhini, S. Sivanandam, M. Rajalakshmi, and M. D. Sidheswaran, "Enhancing the spam email classification accuracy using post processing techniques," *International Journal of Applied Engineering Research*, vol. 10, no. 15, pp. 35125–35130, 2015.

[35] R. T. Pashiri, Y. Rostami, and M. Mahrami, "Spam detection through feature selection using artificial neural network and sine–cosine algorithm," *Mathematical Sciences*, vol. 14, no. 3, pp. 193–199, 2020.

[36] J. Hosseinkhani, M. Nematollahi, M. Akhlaghpour, H. H. Sadegh, Z. S. Arbabi, and E. Ahrari, "Adaptive spam email detection using support vector machines (svms)," 2016.

[37] M. Abdullahi, A. D. Mohammed, S. A. Bashir, and O. O. Abisoye, "A review on machine learning techniques for image based spam emails detection," in *2020 IEEE 2nd International Conference on Cyberspac (CYBER NIGERIA)*, pp. 59–65, IEEE, 2021.

[38] O. Taylor and P. Ezekiel, "A model to detect spam email using support vector classifier and random forest classifier," *Int. J. Comput. Sci. Math. Theory*, vol. 6, pp. 1–11, 2020.

[39] Y.-K. D. Ng and M. S. Pera, "Spamed: A spam email detection approach based on phrase similarity," 2009.

[40] S. Kharazmi and A. F. Nezhad, "Spam email detection using bayesian spanning tree," in *National Conference of Computer Science*, 2007.