MINI PROJECT – II
(2018-19)

# SPAM DETECTION

## System Specifications Requirements


**Team Members**

Kankshit Adhaulia
(161500254)
Deepak Singh
(161500189)
Gaurav Bhardwaj
(161500215)
Mukul Bhardwaj
(161500331)

Supervised By
Mr. Subhash Agrawal
Assosiate Proff.
Department of Computer Engineering and Applications

# TABLE OF CONTENTS

# ABSTRACT

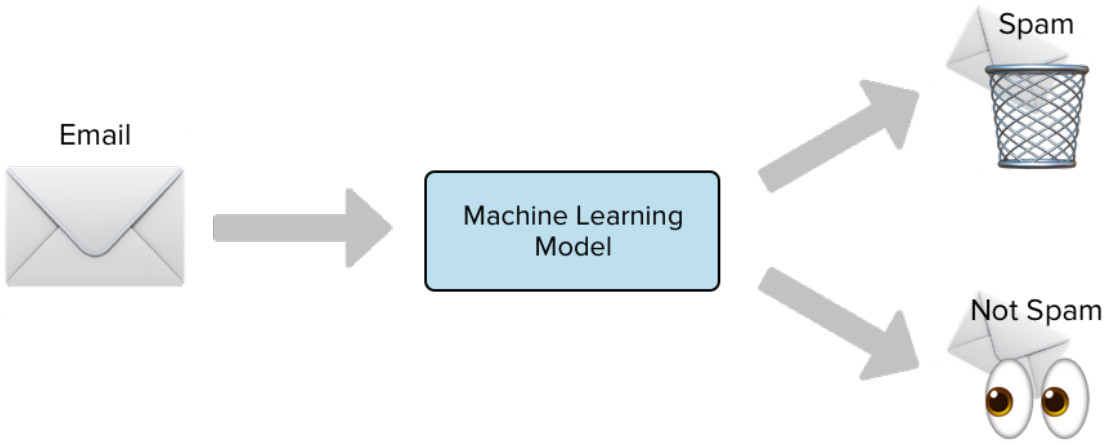
Unsolicited bulk emails, also known as spam emails, are a regular occurrence for anyone who uses email. Spam filtering is a way to distinguish between spam emails and regular emails. The goal with spam filtering is to determine whether an email is spam or not spam, then filtering out the spam emails, resulting in a spam-free in-box for the user.

Logistic regression is a statistical method that can be utilized for spam filtering. It is sensible that spam emails typically share a certain type of characteristics. Words that recurrently show up in spam emails can be used as predictor variables in the logistic regression model. Other email characteristics, such as special formatting, tables, links, may also be used as predictor variables .

This report looks into what determines the probability of an email being a spam email by using logistic regression. We will examine if certain characteristics alter the probability of an email being a spam email or not. We will also test which model best predict the probability of an email being spam probability of an email.

# 1.INTRODUCTION

The idea of this post is to understand step by step working of the spam filter and how it helps in making everyone life easier. Also, next time when you see a “You have won a lottery” email rather than ignoring it, you might prefer to report it as a spam.




The above image gives an overview of spam filtering , plenty of emails arrive everyday, some goes to spam and rest stays in our primary inbox(unless you have further categories defined). The blue box in the middle—Machine Learning Model, how does it decide which mail is spam and which one is not.

Before we start talking about the algorithm and the code, take a step back and try relating that simple explanation of spam detection with monthly active Gmail account(which is approximately 1 billion). The picture seems pretty complicated, isn’t it? Let’s get an overview on how does gmail use the filtering for a huge number of accounts.

## 1.1 Gmail Spam Detection

We all know the data Google has, is not obviously in paper files. They have data centers which maintain the customers data. Before Google/Gmail decides to segregate the emails into spam or not spam category, before it arrives to your mailbox, hundreds of rules apply to those email in the data centers. These rules describe the properties of a spam email. There are common types of spam filters which are used by Gmail/Google.

*Blatant Blocking*- Deletes the emails even before it reaches to the inbox.

*Bulk Email Filter*- This filter helps in filtering the emails that are passed through other categories but are spam.

*Category Filters*- User can define their own rules which will enable the filtering of the messages according to the specific content or the email addresses etc.

*Null Sender Disposition*- Dispose of all messages without an SMTP envelope sender address. Remember when you get an email saying, “Not delivered to xyz address”.

*Null Sender Header Tag Validation*- Validate the messages by checking security digital signature.

## 1.2 Machine learning

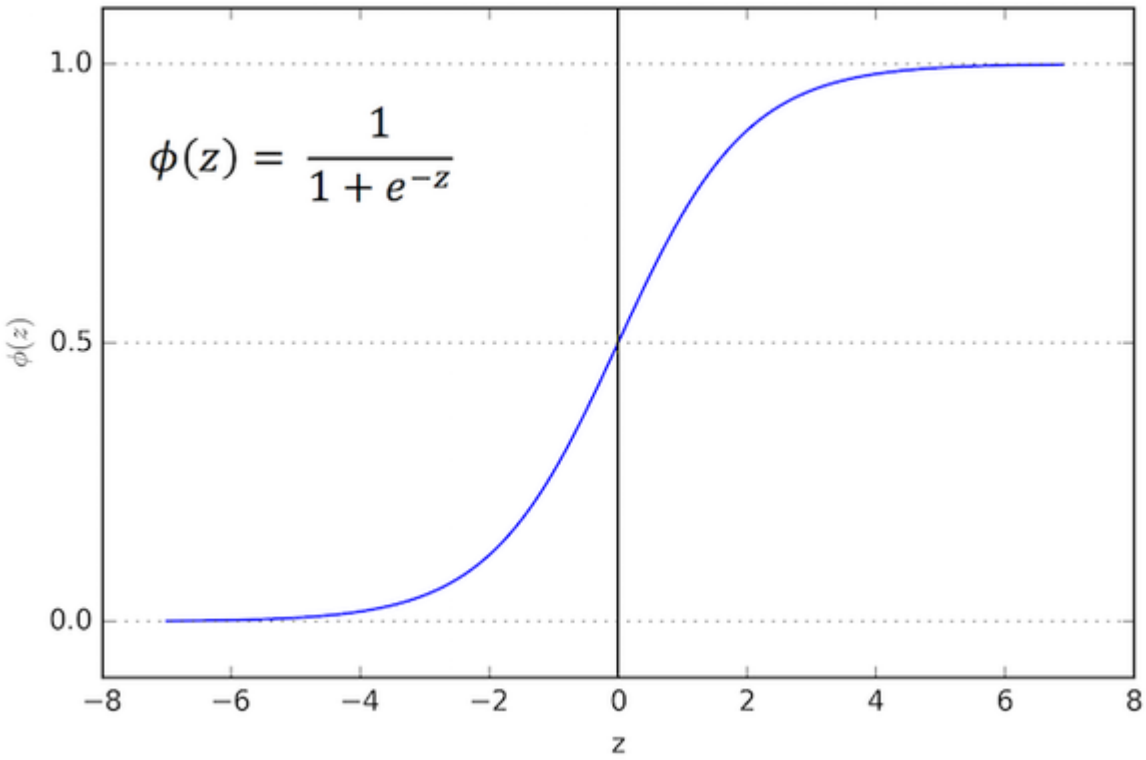
The increasing volume of unsolicited bulk e-mail (also known as spam) has generated a need for reliable anti-spam filters. Machine learning techniques now days used to automatically filter the spam e-mail in avery successful rate. In this paper we review some of the most popular machine learning methods (Bayesian classification, k-NN, ANNs, SVMs, Artificial immune system and Rough sets) and of their applicability to the problem of spam Email classification.

## 1.3 Logistic Regression

According to Wikipedia definition,

> *Logistic Regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function.*

From the definition it seems, the logistic function plays an important role in classification here but we need to understand what is logistic function and how does it help in estimating the probability of being in a class.



The formula mentioned in the above image is known as Logistic function or Sigmoid function and the curve called Sigmoid curve. The Sigmoid function gives an S shaped curve. The output of Sigmoid function tends towards 1 as $z \rightarrow \infty$ and tends towards 0 as $z \rightarrow -\infty$. Hence Sigmoid/logistic function produces the value of dependent variable which will always lie between [0,1] i.e the probability of being in a class.

## 1.4 Types of spam

### 1.4.1 Email Spam

Email spam is the most familiar spam that most of the users come across every day. Email spam follows three properties i.e., anonymity, mass mailing and unsolicited emails. Anonymity is the property of hiding the uniqueness and whereabouts of the email sender. Mass mailing is defined as the sending of bulk identical emails to the large number of groups and unsolicited emails are the emails transferring to the recipients who do not request. Typically, an email sent to large number of groups without any request by hiding their identity is referred as email spam.

### 1.4.2 Comment Spam

This is the most common spam that many users come across in various blogs. Spammers use the posts in the blog to redirect to spam websites. The ranking of such blogs gets increased gradually in the search engines. It is basically used to promote the searching services like Wikipedia, blogs, guest books etc. There are number of tools in the market to get rid of comment spam.

### 1.4.3 Junk fax

Junk faxes are not as prevalent as before. It reduced periodically with the existence of internet technology. However, there are also some risk factors occurring in few corners because of this telemarketing technology. This is similar to junk email where the advertisements and messages are passed to numerous users via fax machines. The adversaries use broadcast fax as a medium to pass on the junk fax to various users. Fortunately, there are surplus tools to overcome junk fax.

### 1.4.4 Unsolicited text messages

This is kind of similar to instant messenger spam but here the messages are passed via mobiles. SMS is the service through which the messages are transferred from one

user to other user. The easiest way is to maintain the contact with the known friends instead of strangers. It is relatively easy to find the source where the message is coming from with the instant messenger spam. It is critically important no to click on the links that are passed via mobile by the spammers.

1.4.5 Social networking spam

Social networking sites play an important role in today's world. With the advent of such sites, spammers also started flooding using new techniques to make the social networking sites such as face book, twitter, linked in etc. as part of the spamming activities. As of now it is targeting only the wall posts, messages but these techniques evolve certainly over a period of time. Spammers use notes or messages through various groups or pass the messages with embedded links, which may lead to pornographic or other sites and target spam . Even though these sites have an option to report spam or abuse activities, the spammers frequently change their address or account to hide their identities.

## 1.5 Problems with spam

1.5.1 Viruses

Viruses are the most dangerous threats across the network. There are many techniques and methodologies developed to decrease the nefarious activities caused by different types of viruses. With the increase in the internet technology, wide variety of viruses produced to attack the machines. Spam is one of the sources to launch such types of viruses. The widely spread viruses are the ones which disconnect the hosts and get diffused into the network. Spam viruses in modern technology are more dangerous as it controls the machine itself and then annihilates them.

1.5.2 Server problems

Most of the time servers are being targeted by the spammers. Due to increase in the intensity and volume of the spam, the company or any system has to use huge resources to maintain the server. In order to distill and disseminate the data that is transferring in

the network more energy costs and resources are to be divided among the departments. Due to this frequency of spam, the performance also gets affected. So, the servers must maintain a low and necessary data. Otherwise, it can create major problems on the server to maintain and causes heavy load disrupting the entire network.

1.5.3 Hacking and Phishing

As the computers in the modern technology are becoming more and more secure, the spammers face more difficulty to capture the confidential details. So, they tend to use various methods to break through the security of different IT departments. Spammers make use of hacking methods like entering into the trusted employee system without the user's awareness. Then, spammers perform different activities and keep a record of the confidential data or hold vital information either for the cost or for self-happiness. Another way is to trap the employees of the companies to enter the passwords or any valuable information into the spammer's website, so that it keeps track of the password to reveal important credentials. Though there are many firewalls and spam filters, spammers are also improving their technical skills to intrude on organizations.

1.5.4 Productivity threats

It is known fact that most number of employees in any organization spends approximately an hour of time to sort out and delete the spam from a cluster of good emails. This leads to heavy wastage of resources like labor cost, time and space in any system. An important email among the cluster of non-spam emails seems to be an unimportant one. This causes problems like loss of e-mail, deleting email and can also disturb the valued customer trust and internal correspondence.

## 1.6 Types of email spam filters

Spam filter is a piece of software that is used to filter the spam emails based on the content and rules adhered by its corresponding software. Every single spam filter has its own set of rules through which the spam is filtered from spreading across the network. It involves the content of the spam, address of the users and where it is redirecting to etc. Based on these parameters it judges, whether an email is a spam or not. There are

multiple spam filters divided based on their rules .

### 1.6.1 Challenge-Response spam filter

This spam filter is a basic filter mechanism that is used to control the spam in the emails. This does not allow any strangers or any pre-approved persons to send an email to the user. In return to the email sent by these pre-approved persons, it asks to validate them in order to pass on the email. The logic behind this strategy is that the pre-approved 6users do not have time to validate their own email ID from thousands of emails that it might have sent. However, there are numerous problems with respect to this system. There are high chances that the spammer uses its fake address in order to validate email address.

### 1.6.2 Rule based scan filtering system

These are referred as the original spam filters. It works on the method of detecting pre-determined words or phrases that most of the spammers use. It identifies those key words and block's emails from passing on from one user to other users. The rules to detect words or phrases are to be improved daily. Because the spammers are so intelligent that it keeps track of words that are blocks and uses its synonyms for a successful transmission. Nevertheless, strict based rules also become another problem, as it blocks even a legitimate email. There is every possibility that both spam and non-spam email get blocked due to the demanding rules passed on by the rule based to scan filtering system. Suppose there is a word spam involved in the message, it gets blocked by the rules based on the filtering system. If the rules are really weak, then spammer' tries to modify the message from "spam" to "sp@m" and passes on the message successfully and spreads the spam across the network. So, there needs to maintain a different strategy for the rule based scans. Even so, one has to accept the fact that the effects of spam can be reduced, but it's impossible to block hundred percent of spam and let the good email pass through the system.

### 1.6.3 Global black lists spam filter

Global black lists contain the list of the notable spammers. Whenever an email is sent, the internet keeps track of the email sender details, i.e., from where the email has

come from. Its IP address also gets recorded. So this spam filter compares against the black list containing the notable spammers. If there is a match, it discards the message before reaching it to the recipient. This makes users to escape from the notable spammers. The black list also gets updated so that it can reduce the well-known spammers to perform the spam activity again. This does not waste its space and time by verifying repeatedly with the spammers but perform only one search with just one database thus saving lot of time and space.

There are certain problems with this filtering system. The black list is decided by the users familiar with detecting spam. So before a spammer getting into this list, there happens to transfer thousands of emails to the users. Sometimes, the legitimate person may get into this black list thus getting boycotted completely without any illegitimate activity. And also, complete internet service providers are blocked because of few users involve in spamming activities. This results in disrupting the entire network.

1.6.4 Bayesian Analysis

All the methods that were discussed above are based on some predictive methods and content in the message during the exchange of information. However, this filtering system is based on the mathematical formulae through which the email can be determined if it is a spam email or non-spam email. As the black lists take a lot of time to get itself updated, the spammers in the mean while pass on the spam across the network. 9So, in this filtering system it just uses a sample of data and determines if the email is a spam or not in short duration.

This is also associated with set of problems. This formula is developed long ago. So, there is every possibility that the spammers get updated and escapes with the sample of data that is used in the formulae. The key point is to know how properly the formula is developed and how well it is reliable. Implementing the same formulae after so many years is also a not good idea.

1.3.5 Classification based spam filter

a spam filter using logistic regression. We will classify messages to be either ham or spam. The dataset we'll use is the SMSSpamCollection dataset. The dataset contains messages, which are either spam or ham.

We load the dataset using pandas. Then we split in a training and test set. We extract text features known as TF-IDF features, because we need to work with numeric vectors.

Then we create the logistic regression object and train it with the data

We can use a language translator to translate text from one language to another. There are various APIs and modules for this, we'll use the Google Translate API.

We will use the Goslate module to translate. Apart from translation, it supports language detection, batch translation, dictionary lookup and more.

# 2. Proposed System

An explanation of logistic regression begins with an explanation of the logistic function(also called the sigmoid.The logistic function is useful because it can take as an input, any value from function).

f (z) =1 + e −z

negative infinity to positive infinity, whereas the output is confined to values between 0 and 1. The variable,z represents the exposure to some set of risk factors, while f(z) represents the probability of a particular outcome, given that set of risk factors. The variable z is a measure of the total contribution of P all the risk factors used in the model and is known as the logit.

Logistic regression is a statistical method used to demonstrate if a binary response variable Y is dependent on one or more independent variables X =(X 1 , · · · , X n ) . It is a tool for for building a model in situations where there is a two-level categorical response variable, in contrast to a numerical response variable, where multiple linear regression would be more appropriate. Like multiple regression, logistic regression is a type of GLM 3 with the difference being the categorical response variable.

The outcome of a GLM is usually denoted by $Y_i$ , where i stands for observation

number i .In this report, $Y_i$ will denote if an email is spam or not; ($Y_i = 1$) for spam,

and ($Y_i = 0$) for non-spam. The independent variables X will take on the following form; $x_{ij}$ denotes the value for variable j for observation number i .The

outcome Y i takes on value ($Y_i = 1$) with probability $\pi_i$ and ($Y_i = 0$) with probability ($1 - \pi_i$) .

The logistic regression model links the probability of an email being spam ($\pi_i$)

to the prediction variables ($x_{1i}$ , · · · , $x_{ij}$) through a framework very similar to

that of multiple regression. Since the response is binary, we need to and a

suitable transformation in order to make the regression model work. A natural

transformation for $\pi_i$ is the logit transformation :

$$\text{logit}(\pi_i) = \ln((\pi_i)/(1 - \pi i)) \qquad .....1$$

The logistic regression model is given by:

$$\ln((\pi_i)/(1-\pi_i)) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} \qquad .....2$$

Note that since the probability of an email being spam ( $\pi$ i ) is a number between zero and one,the log( $\pi$ i/(1−$\pi$i)) can take on any real number:

$$0 \leq \pi_i \leq 1 \Longrightarrow -\infty < \ln(\pi_i/(1-\pi_i)) < +\infty$$

The relation between P (Y i = 1) is obtained by solving 2 for $\pi$ i . We get:

$$P(Y = 1|X = x) = \pi_i = ((\exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij})/$$

$$1 + \exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij})) \quad .....3$$

Equation 3 is the logisic regression model that will be utilized throughout this project.

We define the odds as

$$\Omega = (\pi_i/(1-\pi_i))$$

where the odds is the probability of the outcome spam divided with the probability of the outcome no spam . By taking the logarithm on both sides we get equation 2. The logistic regression coe#cients correspond to the change in the log odds, for each variable respectively. The exponentiated form of the coefficients correspond to the odds ratio.

# 3. Requirement Analysis and Feasibility Study

## 3.1 Requirement Collection and Specification

The Primary data were collected through our supervisor from the related corpus. The secondary data were collected from research papers and some test data were sampled data made by ourselves to check for the expected output.

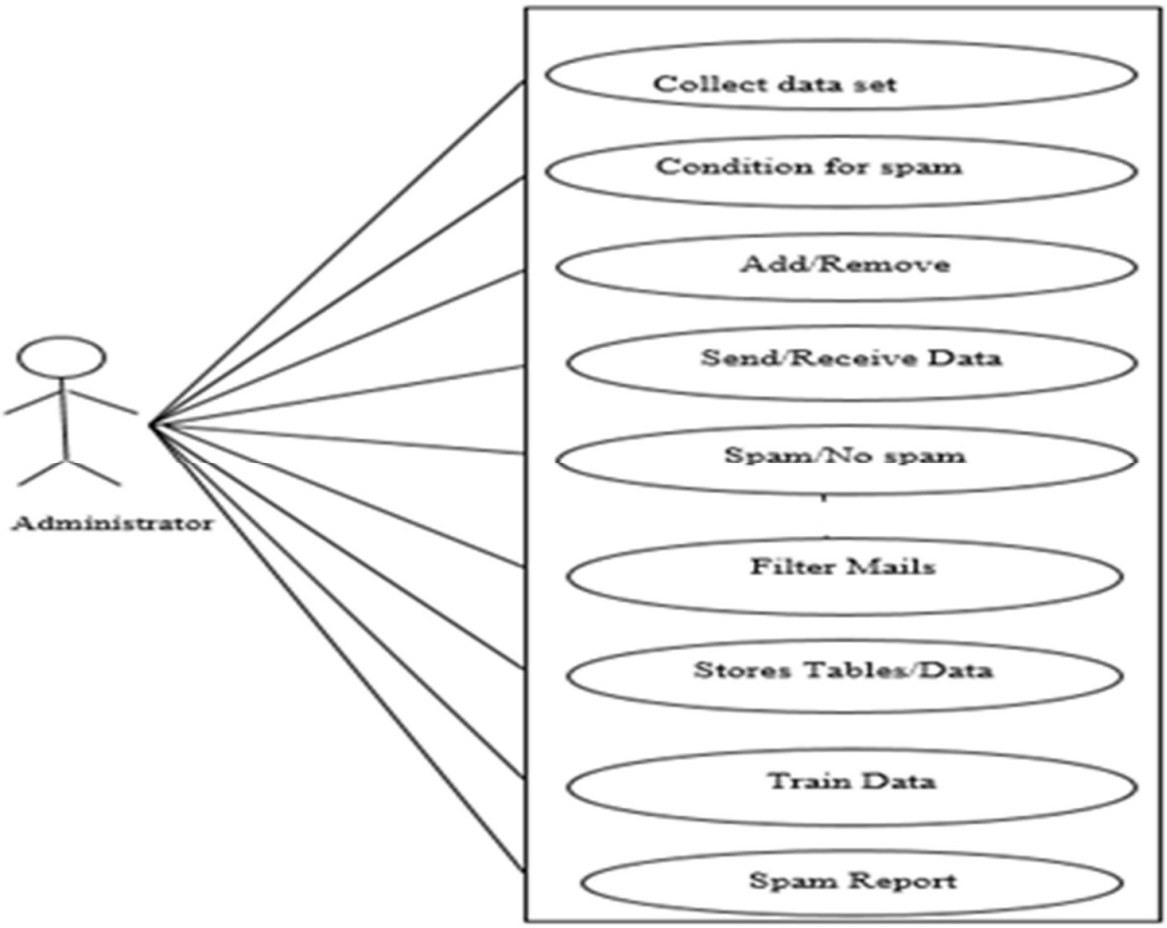
• Corpus is maintained

• The user should be able to receive the genuine mails.

## 3.2 Functional Requirements

The main function of this project is to classify the e-mails which is done by first taking out the feature vector extraction which involves first taking out whether the word is a spam or not by representing in the form of matrix.

## 3.3 Use case diagram




## 3.3 Non Functional Requirements

• Ensures high availability of email data here datasets.

• User should get the result as fast as possible.

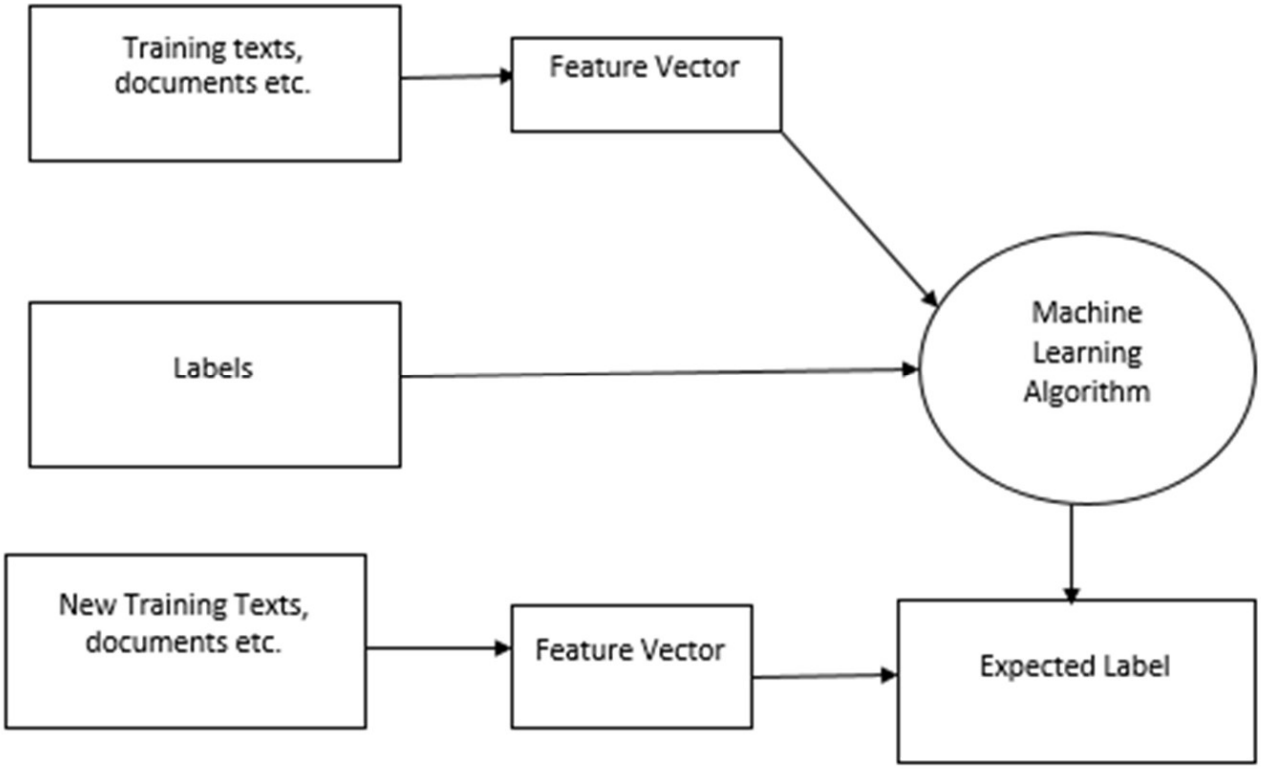
• It should be easy to use i.e., user is just required to type the words and click then the result is displayed

or user is just required to enter a pair of reasonable sentence.

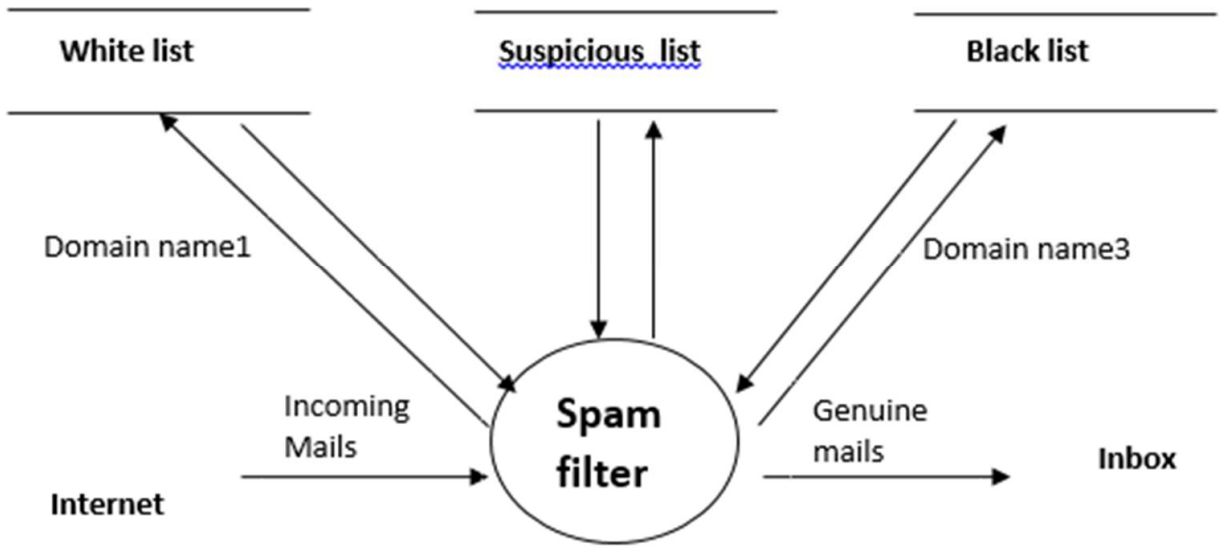
## 3.4 Data Flow Diagram

It is a directed graph where nodes represents processing activity and arc represent data items transmitted between processing nodes.

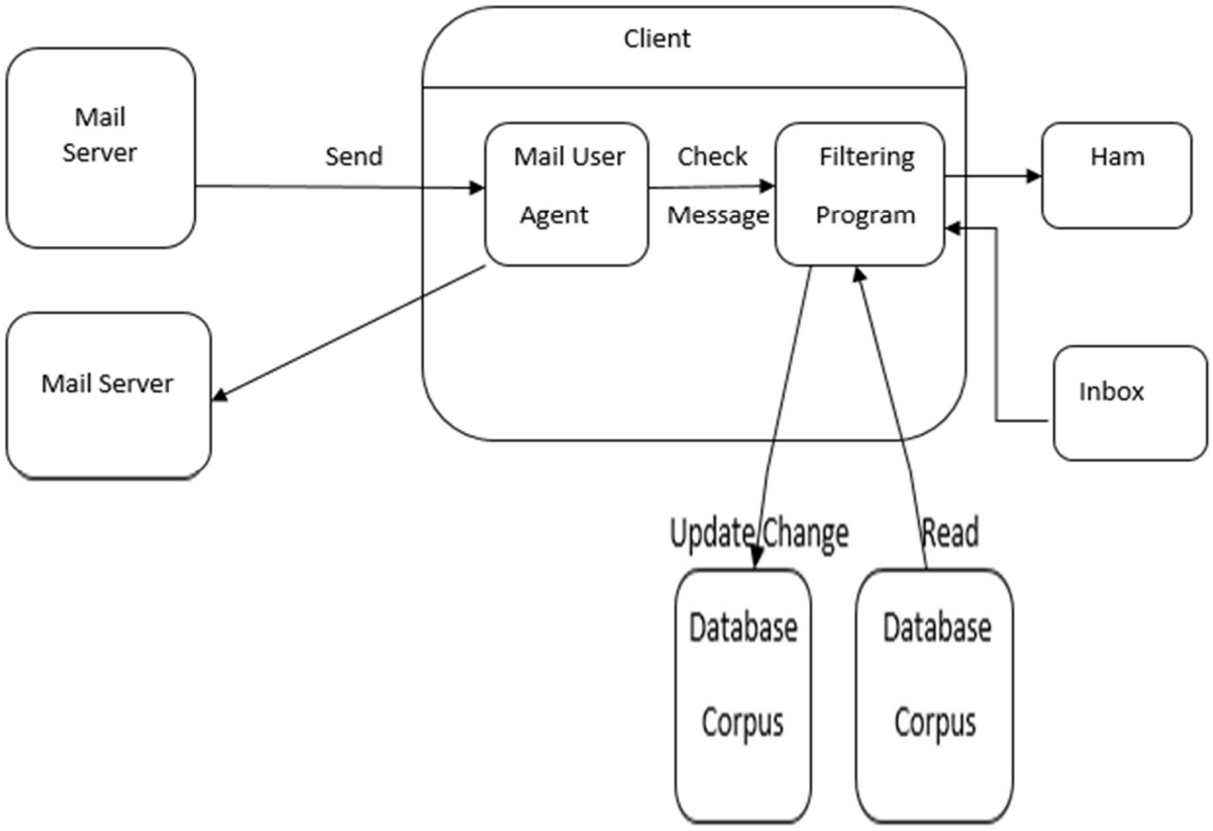
Level 0




Level 1


White list
Suspicious list
Black list
Domain name1
Domain name3
Incoming
Mails
Spam
filter
Genuine
mails
Inbox
Internet

Level-2


Client
Mail
Server
Send
Mail User
Agent
Check
Message
Filtering
Program
Ham
Mail Server
Inbox
Update/Change
Read
Database
Corpus
Database
Corpus

## 3.5 Feasibility Study

The feasibility of the system has been studied from the various aspects like whether the system is feasible technically, operationally and economically. The present technology is found to be sufficient to meet the requirements of the system. This system is believed to work well when it is developed and installed.

Hence, operational feasibility is achieved. Since the requirements for the project are easily available, wareheaded with the intention to use the available resources to fulfill the system requirement.

### 3.5.1 Technical Feasibility

The technology needed for the proposed system that we are going to develop is available. We can work for the project is done with current equipment existing tools like python .We can develop our system still using this technology if needed to upgrade. In future, if we want to use new technology like android app of our system it is possible. Hence, the system that we are going to develop will successfully satisfy the needs of the system for technical feasibility.

### 3.5.2 Economic Feasibility

Since the system is developed as a part of project work, there is no manual cost to spend for the proposed system. Also all the resources are already available, it gives an indication that the system is economically possible for development. Economic justification is generally the "Bottom Line" consideration for most systems. The cost to conduct a full system investigation is negotiable because required information is collected from internet.We can run our system in our normal hardware like desktop, laptop mobiles and so on. This system won't require extra specific software to use it. Hence, the project that we are going to develop won't require enormous amount of Money to be developed so it will be economically feasible.

### 3.5.3 Operational Feasibility

The user interface will be user friendly and no training will be required to use the application. The solution proposed for our project is operationally workable and most likely convenient to solve the irrelevant document and fraud e-mail.

# 4.Result And Analysis

The following table shows the evaluation measures for spam filters.

## 4.1 Evaluation Measure and Evaluation Function

Accuracy Acc = TN+TP/(TP+FN+FP+TN)

Recall R= TP/(TP+FN)

Precision P= TP/TP+FP

F-Measure F= 2PR/(P+R)

(i) Accuracy: Percentage of correctly identified spam and not spam message.

(ii) Recall: Percentage spam message manage to block.

(iii) Precision: Percentage of correct message for spam e-mail.

(iv) F-measure: Weighted average of precision and recall.

## 4.2 Training Evaluation

Accuracy: 96.18%

F1 score: 96.53%

Recall: 96.07%

Precision: 98%

## 4.3 Test Evaluation

Recall= 67.55 %

Accuracy 97.72%

F1 score=70%

Precision=94.61%