A

Major Project Report

On

SPAM DETECTION ON MOBILE PHONE SMS PERFORMACE USING FP-GROWTH AND NAVIE BAYE’S CLASSIFIER

Submitted By:

M. JHANSI PRIYANKA	(18M91A0545)
N. UPENDRA	(18M91A0550)
R. LASYA PRIYA	(18M91A0556)
L. NIHARIKA	(19M95A0503)

DEPARTMENT
OF
COMPUTER SCIENCE AND ENGINEERING

AURORA’S SCIENTIFIC AND TECHNOLOGICAL INSTITUTE
Approved by AICTE, affiliated JNTU Hyderabad
GHATKESAR -501301, TELANGANA
[2021-2022]

AURORA’S SCIENTIFIC AND TECHNOLOGICAL INSTITUTE

Approved by AICTE, affiliated JNTU Hyderabad
GHATKESAR -501301, TELANGANA
[2021-2022]

DEPARTMENT
OF
COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that Major Project work entitled “ **SPAM DETECTION ON MOBILE PHONE SMS PERFORMANCE USING FP-GROWTH AND NAVIE BAYE’S CLASSIFIER ”** is a bonafide work carried out in the IV/I semester by “ **M. JHANSI PRIYANKA 18M91A0545, N. UPENDRA 18M91A0550, R. LASYA PRIYA 18M91A0556 & L. NIHARIKA 19M95A0503” ”** in partial fulfillment for the award of Bachelor of Technology in Computer science and Engineering from Jawaharlal Nehru Technological University Hyderabad.

Guide **Head of Department**

DR. M. SRIDHAR **DR. M. SRIDHAR**

External Examiner

ACKNOWLEDGEMENT

The completion of this Major Project work gives me an opportunity to convey my gratitude to all those who helped me to complete the Major project successfully.

First, I gratefully acknowledge my deep sense of gratitude to Almighty for spiritual Guidance blessings shown to complete the Major project. I thank my parents for unconditional support to improve myself throughout my life.

My sincere thanks to **Management**, of Aurora's Scientific and Technological Institute, for providing this opportunity to carry out the Major project in the institution.

I owe my respectable thanks to **Dr. R. Mahesh Prabhu (Principal)** of Aurora's Scientific and Technological Institute, for providing all necessary facilities and encouraging words for completion of this Major Project.

I gratefully acknowledge **Dr. M. Sridhar**, **(Head of Department)** Computer Science and Engineering, for his encouragement and advice during this Major project.

My sincere thanks to **Dr. M. Sridhar (Major Project Coordinator)** Aurora's Scientific and Technological Institute, for continuous support for doing this Major project.

I would like to express my thanks to all the faculty members of Department of Computer Science and Engineering, and Non-teaching staff of Aurora's Scientific and Technological Institute who have rendered valuable help in making this Major project successful.

M. Jhansi Priyanka(18M91A0545)
N. Upendra(18M91A0550)
R. Lasya Priya(18M91A0556)
L. Niharika(19M95A0503)

INDEX

FIGURE INDEX

Abstract

SMS (Short Message Service) is still the primary choice as a communication medium even though nowadays mobile phone is growing with a variety of communication media messenger applications. However, nowadays along with the SMS tariff reduction leads to the increase of SMS spam, as used by some people as an alternative to advertise and fraud. Therefore, it becomes an important issue as it can bug and harm the users and one of its solutions is with automatic SMS spam filtering.

One of most challenging in SMS spam filtering is its accuracy. In this research we proposed to enhance SMS spam filtering performance by combining two of data mining task association and classification. FP-growth in association is utilized for mining frequent pattern on SMS and Naive Bayes Classifier is used to classify whether SMS is spam or ham. Training data was using SMS spam collection from previous research. The result of using collaboration of Naive Bayes and FP-Growth performs the highest average accuracy of 90%. FP-Growth for dataset SMS Spam Collection and improves the precision score; thus, the classification result is more accurate.

1. INTRODUCTION

- A **Spam SMS** is any unrequested communication, often sent en masse via the internet or an electronic messaging service.

- Spam also called as Unsolicited Commercial Email (UCE). Involves sending messages to numerous recipient at the same time (Mass Messaging).

- Short Message Service is measured as most extensively used message facility. It is a technique of sending short text messages from one device to another. The usage of mobiles is growing everyday as they deliver a huge variation of facilities by dropping the rate of amenities.

- Due to abundant usage of these services, it has led to growth in mobile device outbreaks like SMS junk. Generally, the word spam refers to the message which is unsolicited. Simply we can state that spam is a junk text message sent from one device to another in the SMS text format.

- On Mobile, the most obvious types of unwanted text messages are un recognized numbers and robo-texts sent by auto-dialers, often promoting a product or service.

- 80% of all spam is sent by less than 200 spammers. Spam messages can take other forms, including phone calls from spam phone numbers, not to mention all the spam emails that inundate our inboxes daily.

Fig 1.1 : Spam Messages in SMS

- These spam messages can cause threat to personal data stored in the device. By the enormous growth in population and increase of all these technological aspects have been growing extremely which in turn increases unanimous spreading of such threat due to less effective security control measure and in order to solve such problems many researchers have developed many techniques to solve and protect the devices from such threats in many different ways.

- The main motive is to provide privacy, convenience and harmony. The classifier used to build these models are Support Vector Machine and Naïve Bayes Classifier, these the two mostly used traditional classifiers. And by convolutional Neural Networks also we can achieve the required model.

- Essentially while training the model firstly, we want to consider a data set, here SMS Spam Collection dataset is used and it is divided into train and test data set and the Naïve Bayes Classifier is used in training and evaluation of the model.

Fig 1.2 : Spam Filter

2. PROJECT ANALYSIS

2.1 Existing System

- Now a day's true caller is that existing system which can block these senders message whose messages are annoying you but we have a control over the sender but not ones the type of messages.
- So, we need such technology/system which can block the particular kind of messages.

Dis-Advantages:

- Suppose it a user don't want any promotional message and it he knows which all users can send him these kinds of messages then he/she can block these senders.
- But if these blocked users any informational message to the user then user will not be able to receive the message.

2.2 Proposed System

- We are using Machine Learning algorithm (Naïve Bayes Algorithm) to eradicate such problem. In this algorithm model will train the machine by its 70% and 30% of dataset.
- Through this 70% data our machine will be trained enough to decide which is the SPAM message or which is the HAM message.

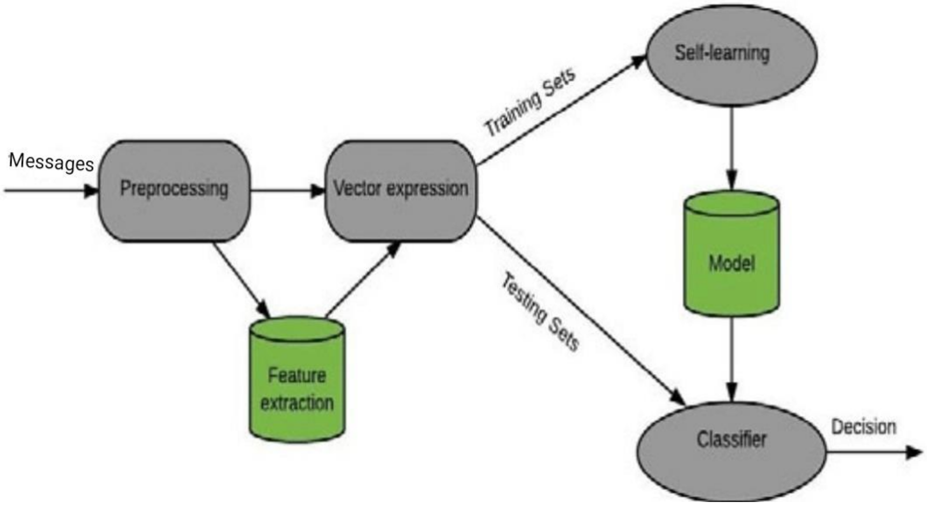


Fig 2.2 : Proposed System For Spam Detection

Advantages:

- we can easily block the unnecessary messages compare to existing system. Then the proposed system will distinguish between SPAM & HAM.
- we are not supposed to block the users we can just oppose or block that type of least important message without blocking the user. So, the users can send any important message.

3. SYSTEM REQUIREMENTS

3.1 Hardware Requirements

- System : Intel CORE i3
- Hard disk : 40 GB
- Floppy Drive : 1.44Mb
- Monitor : 15VGA Colour
- Mouse : Logitech
- Ram : 2 GB

3.2 Software Requirements

- Back End: Machine Learning, Data Set and Algorithms like Navies Baye’s, FP- Growth.
- Technical Tools:
  - Pycharm
  - Jupyter
  - Google
  - Colab

4. SYSTEM ARCHITECTURE

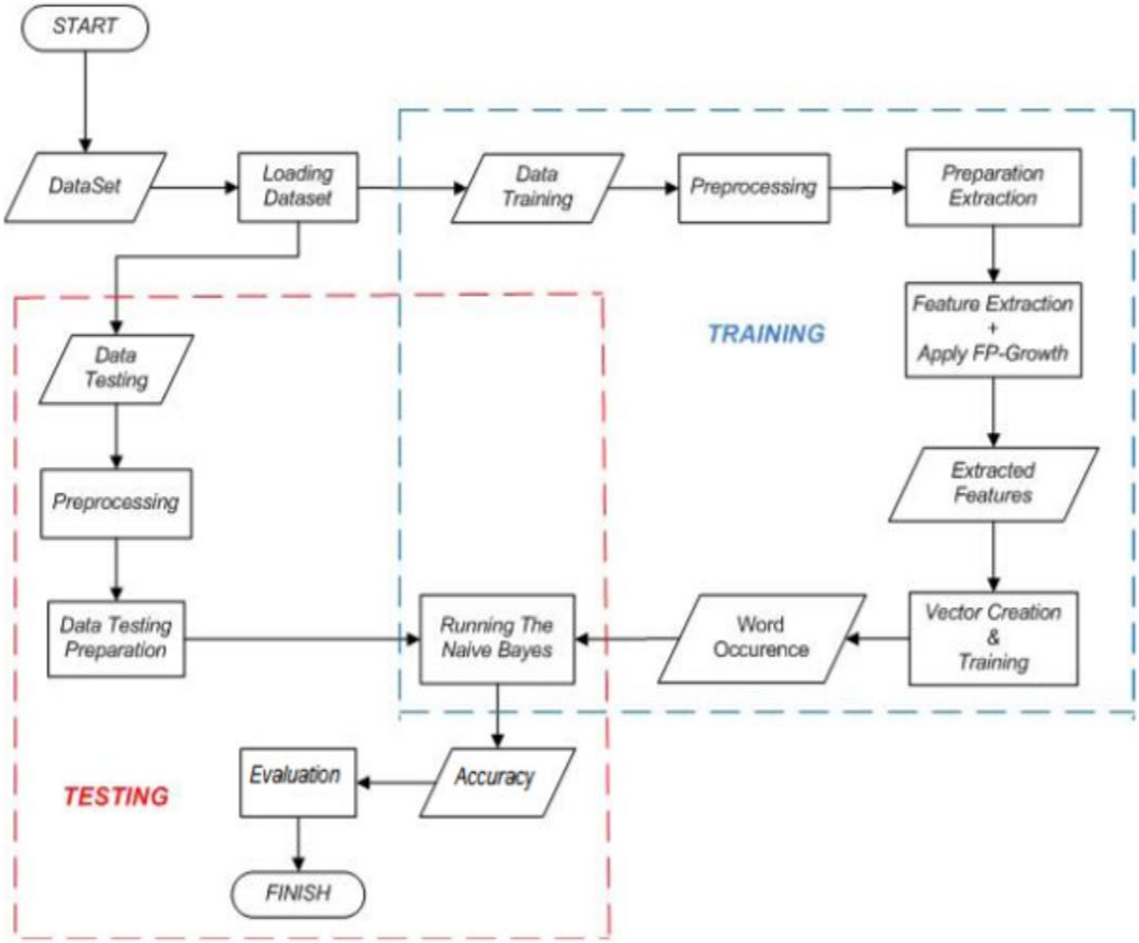


Fig 4 : System Architecture

5. FLOW CHART

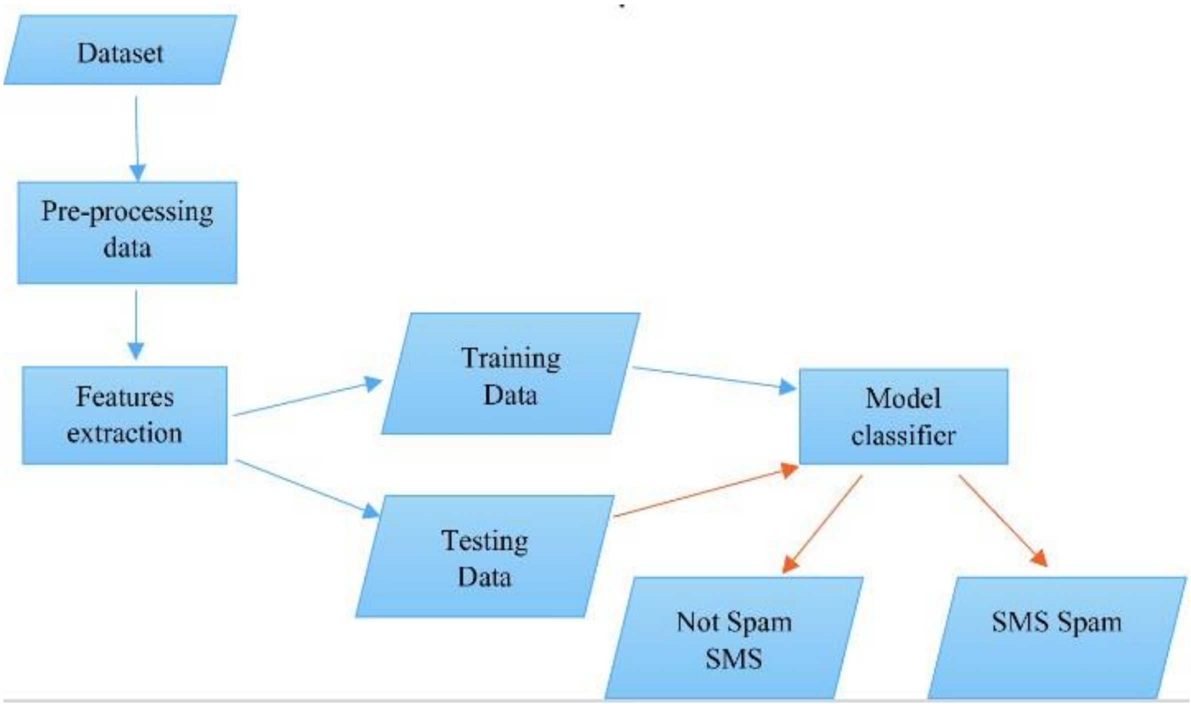


Fig 5 : Flow Chart of Spam Detection

6. MODULES

1. DATA SET AND RUNNING ENVIRONMENT:

- We have manually collected a British English data set consisting of 425 SMS spam messages from the Grumble Text website.
- The Grumble Text website receives SMS spam reported by volunteer users. However, because of privacy concerns, private data, such as name, address and phone number, have been removed.
- The SMS are not chronologically sorted. We argue that our data set is reliable for this research. Compared with another public data set [25],** our data set's messages are in the same language (British English) and from the same society (Britain).

Items	Labels
WINNER! As a valued network customer you have been selected to receives 79000 price. Hurry up!	Spam
A Constant Learner	Ham
Free Entry in 2 a weekly comp to win FA Cup FA Cup final Tickets 21st Mat 2005. Text FA to Book!	Spam

Table 1 : Data Set for Spam Detection

2. FEATURE EXTRACTION:

- An English stop word is used to remove meaningless words because these words exist in both e-mail spam and e-mail ham.
- However, removing such words will reduce the number of useful features to analyze because SMS is very short.
- As a simple example, we have two SMS spam: "$$$ buy free Viagra!" and "free SMS!", also one SMS ham: "buy book :p". We obtain five words: "buy", "free", "Viagra", "SMS", and "book". These words then enter the word vocabulary (|v|).

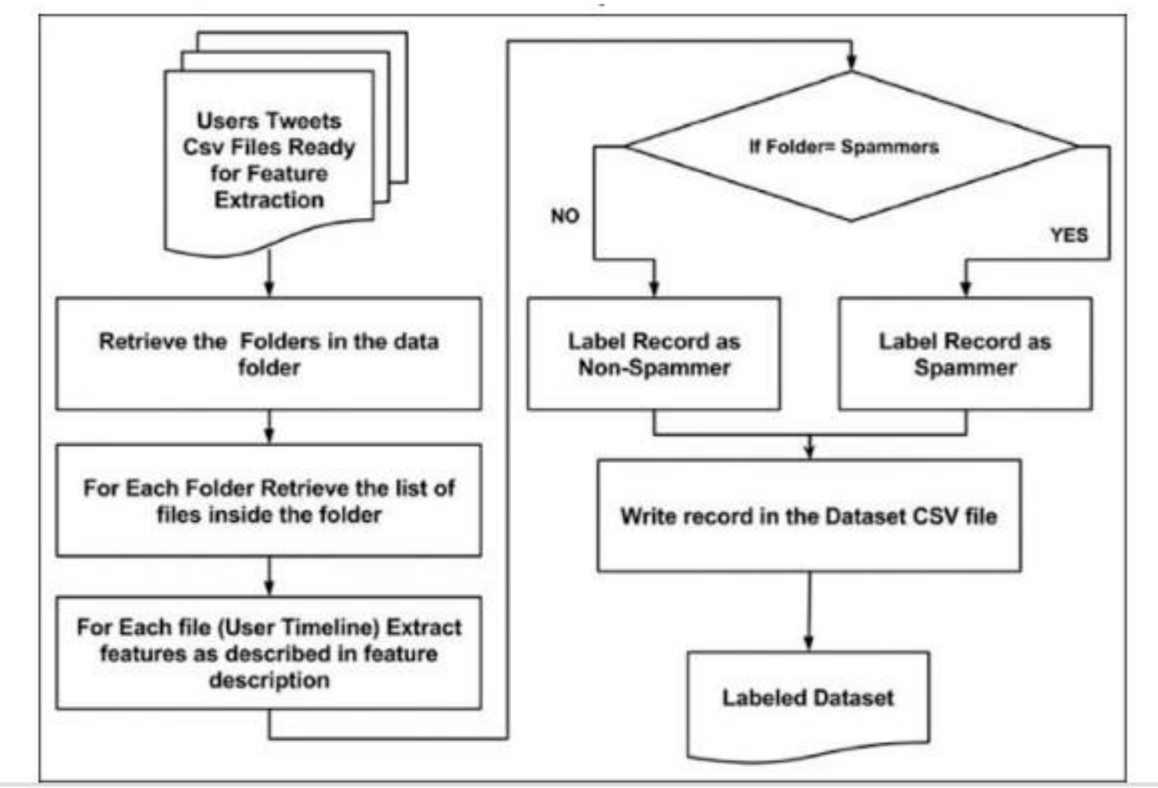

Users Tweets Csv Files Ready for Feature Extraction
Retrieve the Folders in the data folder
For Each Folder Retrieve the list of files inside the folder
For Each file (User Timeline) Extract features as described in feature description
If Folder= Spammers
NO
YES
Label Record as Non-Spammer
Label Record as Spammer
Write record in the Dataset CSV file
Labeled Dataset

Fig 6.2 : Feature Extraction

3. VECTOR CREATION:

- Vector creation is a process used to map a raw tokenized word into numerical data that is ready to be classified by the text classification algorithm.

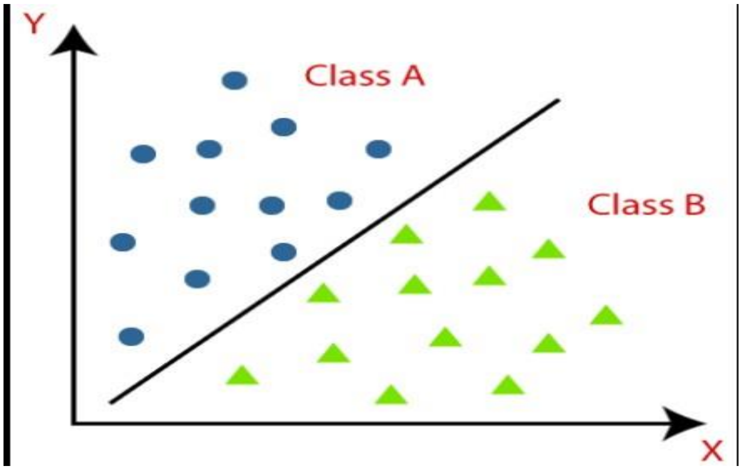


Fig 6.3 : Classification Using Navies Bayes Alogrithm

- We propose to use Word Occurrences because of its simplicity as our proposed approach was designed for mobile phones. We just need to count the number of words in each SMS message.

4. NAIVE BAYES CLASSIFIER:

- Once the word occurrences table is built, we can apply the Naïve Bayes approach to filter unknown incoming SMS. e the probability of SMS ham is higher than the probability of SMS spam, we can say that the SMS is ham.
- Intuitively, a human will agree that the SMS is ham, not spam, because there is no word, such as “free” or “Viagra”, present that is usually present in SMS spam.

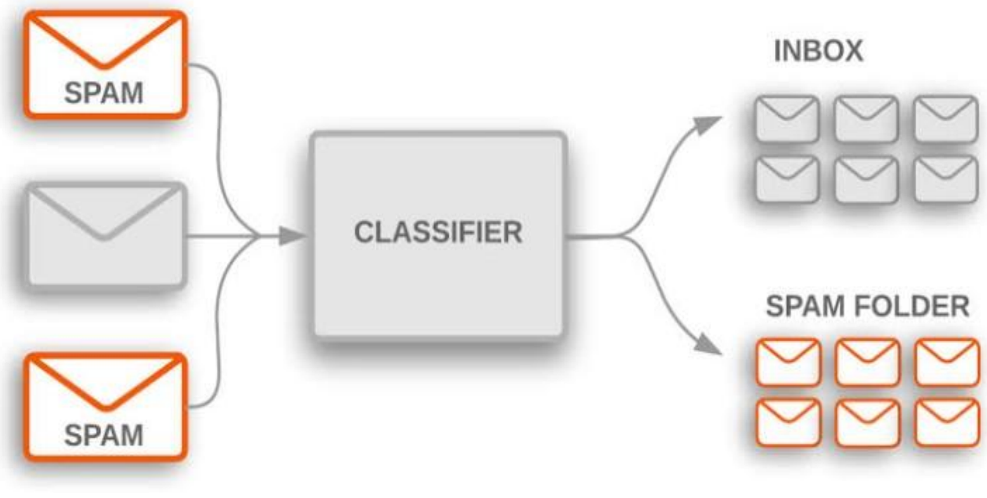


Fig 6.4 : Navie Baye’s Classifier

- However, if the vocabulary count is high and the number of words is high, then the probability value will be too low for the processor to run the   mathematical calculation, especially in a mobile phone. This problem is termed the underflow problem.

5. UPDATING FILTERING METHOD:

- The feature extraction and vector creation steps are part of the training process. The filtering process step is part of the classification process.
- If we receive a new SMS and want to update the filtering system, we just need to repeat the feature extraction and vector creation steps.

- If the word already exists in the word occurrences table, we will just update the word occurrences table. If the word does not exist.

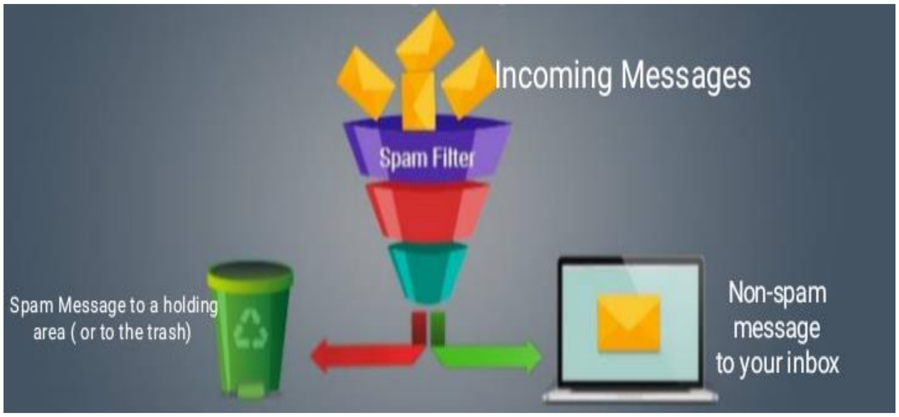
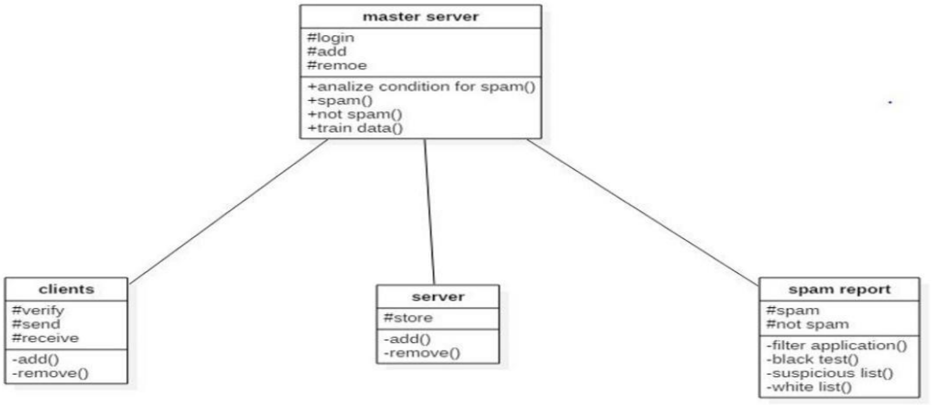

Incoming Messages
Spam Filter
Spam Message to a holding area (or to the trash)
Non-spam message to your inbox

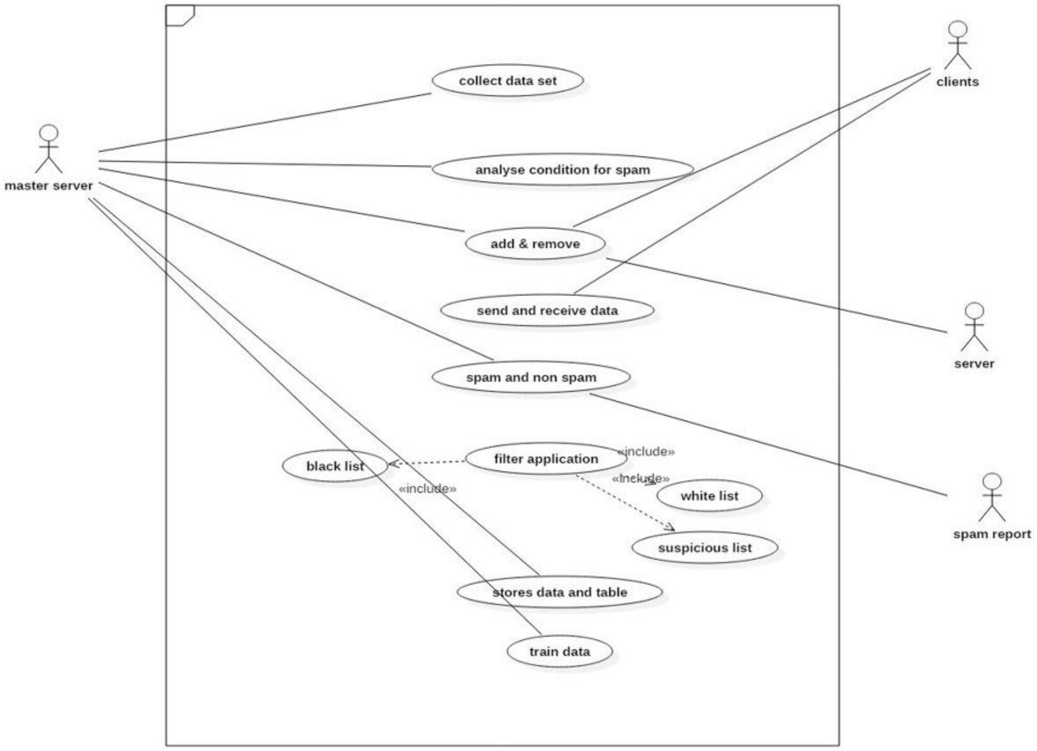
Fig 6.5 : Filtering Method

7. UML DIAGRAMS

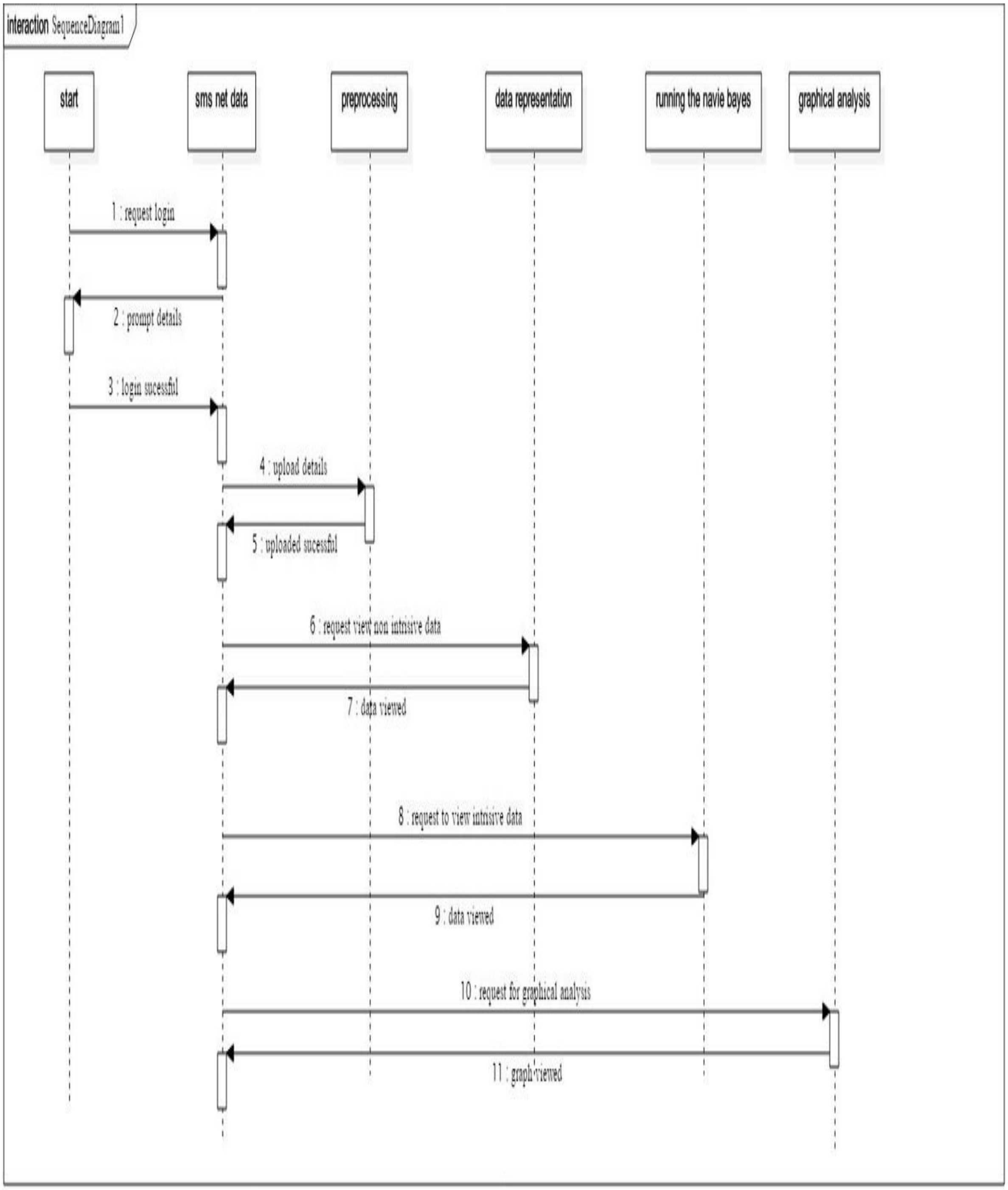
7.1 Class Diagram



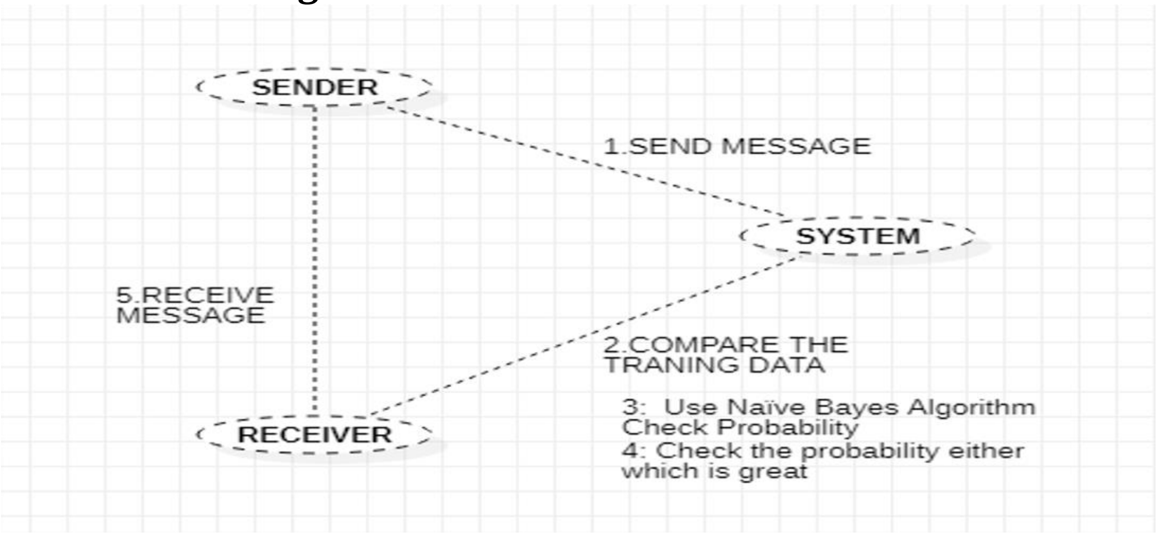
7.2 Use Case Diagram



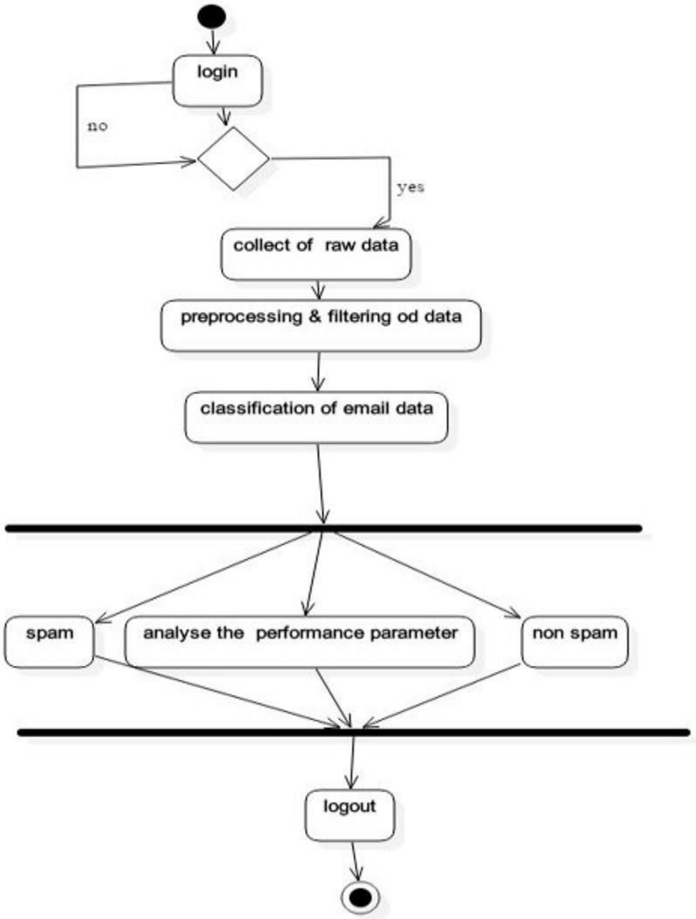
7.3 Sequence Diagram



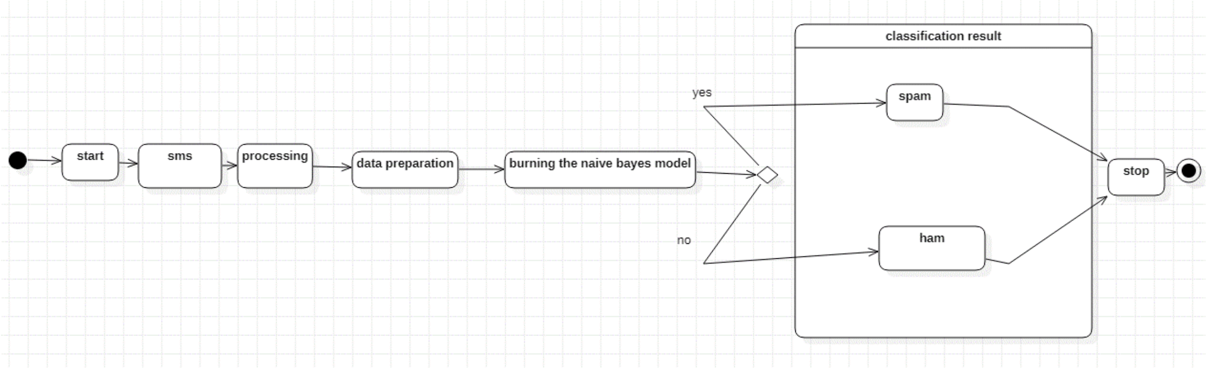
7.4 Collaboration Diagram


SENDER
1.SEND MESSAGE
SYSTEM
5.RECEIVE MESSAGE
2.COMPARE THE TRANING DATA
3: Use Naïve Bayes Algorithm Check Probability
4: Check the probability either which is great
RECEIVER

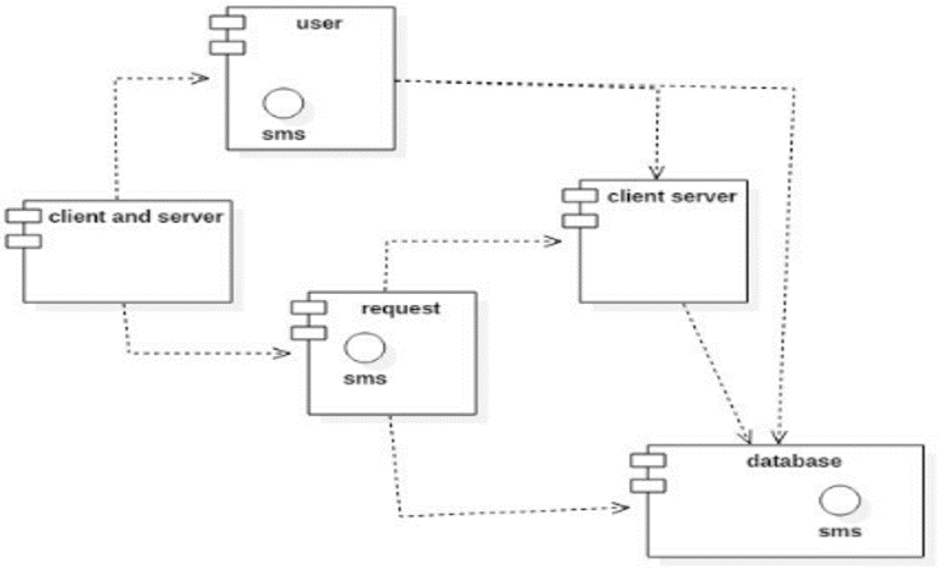
7.5 Activity Diagram


login
no
yes
collect of raw data
preprocessing & filtering od data
classification of email data
spam
analyse the performance parameter
non spam
logout

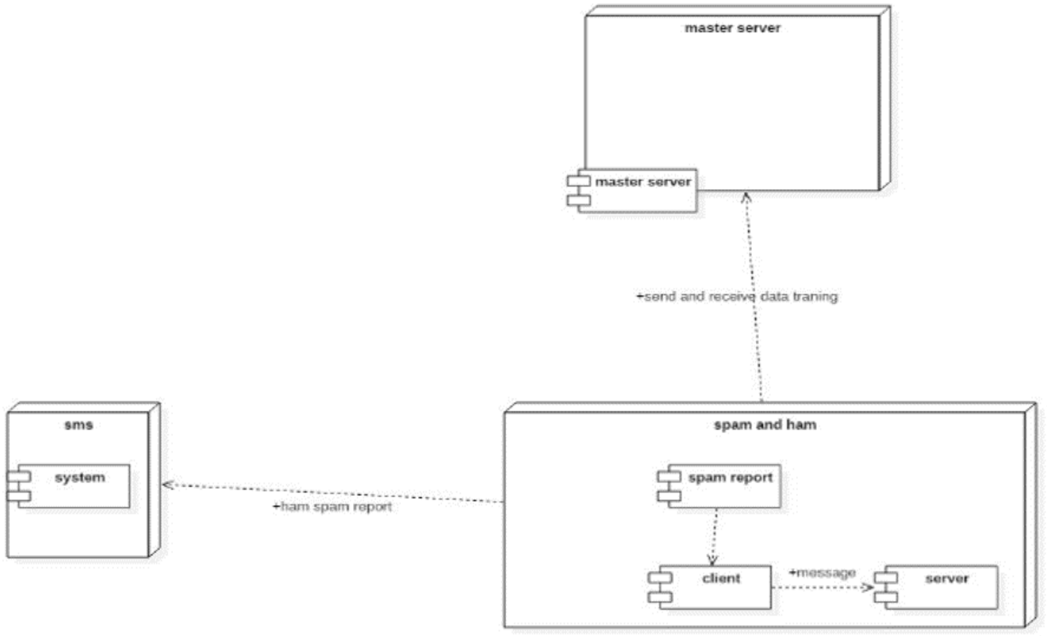
7.6 State Chart Diagram



7.7 Component Diagram



7.8 Deployment Diagram



8. WORKING THEORY OF SPAM DETECTION

Naive Bayes Classifier is one of the simple and most effective classification algorithms which helps in building the fast machine learning models that can make quick predictions.It is a probablistic classifier,which means,it predicts on the basis of the probability of an object.

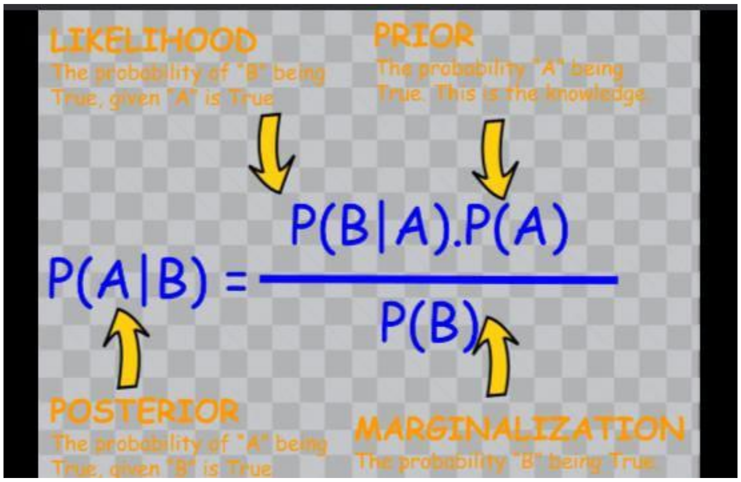

LIKELIHOOD
The probability of "B" being True, given "A" is True
PRIOR
The probability "A" being True. This is the knowledge
P(A|B) = P(B|A).P(A) / P(B)
POSTERIOR
The probability of "A" being True, given "B" is True
MARGINALIZATION
The probability "B" being True

Fig 8.1 : Probability Formulae

P(A/B)=Probability of hypothesis A on the observed event B.
P(A)=Probability of hypothesis before observing the evidence.
P(B/A)=Probability of the evidence given that the probability of a hypothesis is true.
P(B)= Probability of Evidence.

- A Navie Baye's Classifier is an algorithm that uses Baye's theorem to classify objects.
- Navie Bayes is called navie because it assumes that each input variable is independent.
- This is a strong Assumption and unrealistic for real data. However, the technique is very effective on a large range of complex problems.
- Popular uses of Navie Bayes Classifiers include spam filters, text analysis and medical diagnosis.
- For eg: An collection of Shapes.

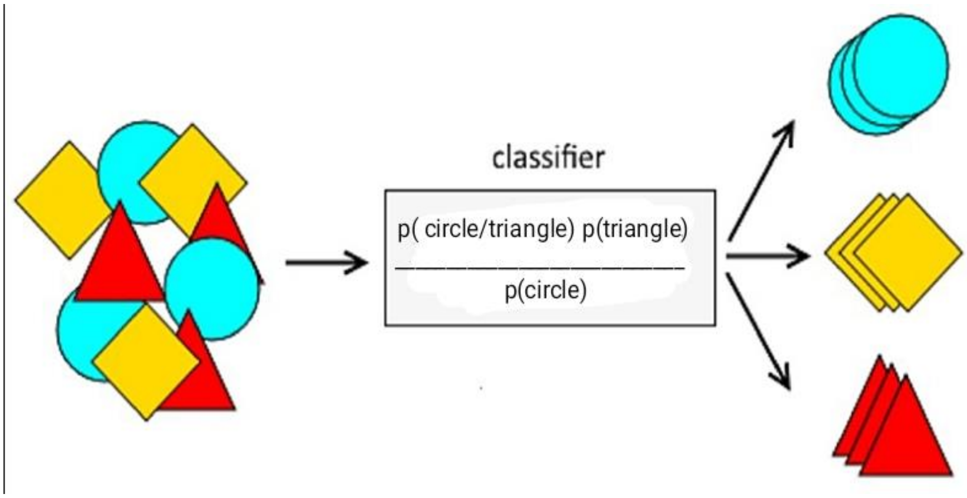


Fig 8.2 : Classifier Using Probability Theory

Algorithm For SMS Spam Classification

Step 1: Retrieving the data

1a. Gather the data from data Sources to retrieve the input data.

1b. Perform Tokenization on input data to form Term Document Matrix.

Step 2: Preparing the data

2a. Apply pre-processing techniques-Stopping, Stemming and Homoglyphing in SMS.

2b. Rank the attributes according to its frequency.

Step 3: Feature selection

Apply one of feature selection technique (Chi-Squarecand Information gain) to select the attributes.

Step 4: Classification

Apply classification technique (Navie Bayes) to classify messages into spam and Non-spam(ham).

Step 5: Evaluation of Result

Evaluate the results using performance parameters like Accuracy, Precision, Recall and F-measure.

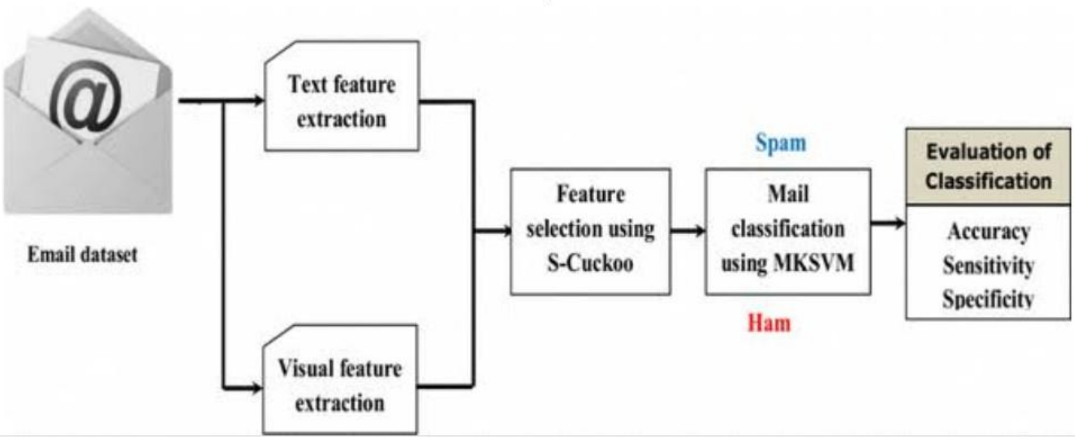

Email dataset
Text feature extraction
Visual feature extraction
Feature selection using S-Cuckoo
Spam
Mail classification using MKSVM
Ham
Evaluation of Classification
Accuracy
Sensitivity
Specificity

Fig 8.3 : SMS and Email Evaluation of Spam And Ham

Accuracy:

The most common definition of accuracy used in marketing anti-spam products is the total number of correctly identified messages divided by the total number of messages. Formally, that is: (s + h) / (S + H) or, in this case **99.27%**. 99.27% sounds pretty good when marketing, but this figure is meaningless.

Accuracy=(s + h) / (S + H)

Precision:

Precision is the fraction of results classified as positive, which are indeed positive. Recall is the fraction of all positive results which were detected. My purpose is to reduce the number of Normal accounts which is labelled as "Spam". This means you want to maximize the precision of Spam and recall of Not spam.

Precision= True Positive/True Positive + False Positive

Recall:

Precision is the fraction of results classified as positive, which are indeed positive. Recall is the fraction of all positive results which were detected.

Recall=True Positive/True Positive + False Negative

F1 Score:

Our spam filter with 0.93 precision and 0.88 recall has an F1 score of **0.90** . F1 is one when a classifier has perfect precision and recall, and goes to zero for classifiers which have either low precision or recall (or both).

F1=2*(Precision*Recall)/(Precision+Recall)

9. DESIGN

Outline :

The Main Goal Of these SMS Spam Detection using FP-Growth and Navies Baye's classifier and as same as modules:

1. EDA (Exploratory data analysis):

   To Investigates Data set and Summarize their main Characteristics, Often Employing data visualization methods.

2. Data Preprocessing:

   Data Preprocessing is a process of preparing the raw data and making it suitable for a machine model. It is the first and crucial step while creating a machine learning model.

3. Feature Extraction:

   It is a general methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient Accuracy.

4. Scoring &Metrics:

   Scoring is widely the process of generating new values, given a model and some new input. The generic term 'score' is used, rather than 'prediction' because scoring process can generate to many different types of values.

   Metrics are Accuracy, Precision, Recall, F1-Score, etc…

5. Flow Analysis:

   The Resulted data can be displayed in form of line charts, Bar Graphs, Pie Charts, Flow Chart,etc…

10. ADVANTAGES AND DISADVANTAGES

Advantages:

- It is easy and fast to predict the class of the test data set. FP-Growth in Association is utilized for mining frequent pattern on SMS and Navie Baye’s classifier is used to classify whether SMS is Spam or Ham.
- Training data was using SMS Spam collection from Previous Research. The result of using Collaboration of Navie Bayes and FP-Growth performs the highest Average Accuracy.

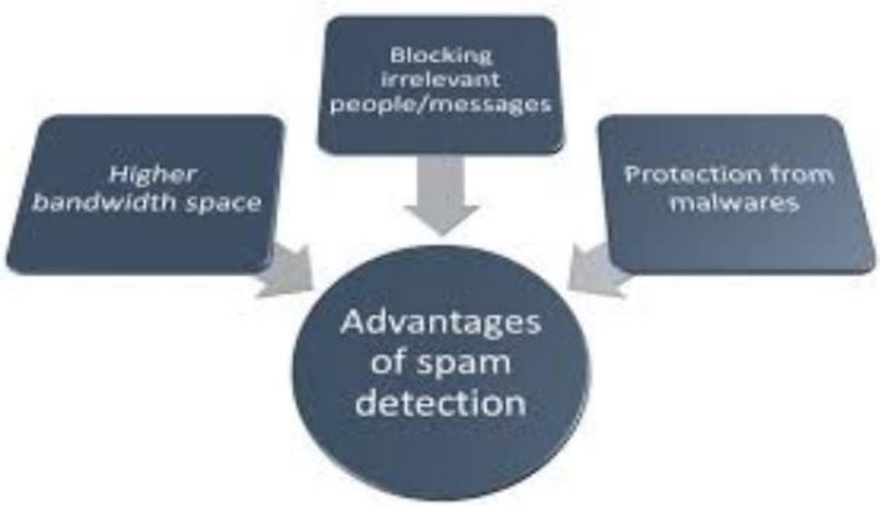

Blocking irrelevant people/messages
Higher bandwidth space
Protection from malwares
Advantages of spam detection

Fig 10.1 : Advantages of Spam Detection

Disadvantages:

- A subtle issue with Navie Bayes Classifier is that if you have no occurrences of a class label and a certain attribute value together then the frequency-based probability estimation will be ‘zero’.
- A big data set is required for making reliable predictions of the probability of each class.

11. CONCLUSION

Based on the analysis of the tests performed in this research, it can be concluded that:

- Both methods used in this research, the performances of both methods is equally well for SMS classification with average of the accuracy above 90%.
- The use of collaboration methods, Naive Bayes and FP-Growth, is superior to the average accuracy for each dataset.
- The Accuracy best average is obtained when the SMS Spam Collection v.1 dataset with the 9%minimum support is used and the implementation of the FP-Growth has accuracy up to 98.506%.
- The use of datasets with varied training data is agreeable to be applied by using the FP-Growth. By implementing the FP-Growth for feature extraction, it can elevate the score of precision.
- Thus, the system becomes more precise in providing the information requested by the users in response to the SMS classification.

12. BIBLIOGRAPHY

1. https://ieeexplore.ieee.org/abstract/document/7811442/

2. https://onlinelibrary.wiley.com/doi/full/10.1002/sec.577

3 https://github.com

4 https://www.grumbletext.co.uk/

5 https://www.dt.fee.unicamp.br/-tiago/smsspamcollection/

6 https://www.esp.uem.es/jmgomez/smsspamcorpus/