

Department of Masters Of Computer Applications

**A Project Report
On**

**Spam detection on mobile phone Short Message Service (SMS)
performance using FP-growth and Naive Bayes Classifier**

Under the Guidance of

Prof: SreeRanjan

**Project By
G Madhu Chandana**

**Spam detection on mobile phone
Short Message Service (SMS) performance using FP-growth and
Naive Bayes Classifier**

BY
Chandana

Agenda

- Abstract
- Problem Definition
- Existing System
- Proposed System
- System Requirements
- Module Description
- System Architecture
- UML Diagrams
- Result
- Conclusion

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.

Problem Definition:

SMS is a text-based communication media that allows mobile phone users to share a short text. Along with the widespread use and popularity as the most important communications media, there are plenty of those who use it for commercial purposes such as advertising media and even fraud. The reduced SMS rate is one of the causes of increasing SMS spam. When we receive any SMS/Mails it may be either HAM (Important messages) or SPAM (least important messages). But sometimes SPAM messages divert our mind. So we need such system which can either block the SPAM message as either it move the SPAM message into different folder without disturbing USER.

What is Spam?

- ▶ Spam is flooding the Internet with many copies of the same message, in an attempt to force the message on people who would not otherwise choose to receive it. Most spam is commercial advertising, often for dubious products, get-rich-quick schemes, or quasi-legal services. Spam costs the sender very little to send -- most of the costs are paid for by the recipient or the carriers rather than by the sender.

Examples of spam/ham:

HAM: Hi this is Saravana how are you abhi.....

HAM: I'm going to try for 2 months ha ha only joking

SPAM: Free Entry in 2K weekly comes to win 2L for monthly enroll
your name.

SPAM: 07732584351 - Rodger Burns - MSG = We tried to call you re your reply to
our sms for a free nokia mobile + free camcorder.
Please call now 08000930705 for delivery tomorrow

HAM: Tomorrow have review to me in my college I am coming to tirupati.

HAM: ok come let's meet here bye....

SPAM:SMS. ac Sptv: The New Jersey Devils and the Detroit Red Wings play Ice
Hockey. Correct or Incorrect? End? Reply END SPTV



EXAMPLE OF SPAM

Weather Report Guy

► Content in Image

Weather, Sunny, High
82, Low 81, Favorite...

**100's of Lenders
Compete for your
Loan to get you
the *Lowest Rate!***

- Refinancing
- New Home Loans
- Debt Consolidation
- Debt Consultation
- Auto Loans
- Credit Cards
- Student Loans
- Second Mortgage
- Home Equity

***Good Credit - Bad Credit
Bankruptcy - Foreclosure***



Interest Rates are at their lowest point in 40 years! We help you find the best rate for your situation by matching your needs with hundreds of lenders!

100% Free Service!

[Click Here To Begin](#)

hawnmyl info
iroqto jn kigu.

Weather

NA, NA - Sunny

High: 82 , Low: 81 degrees

[Favorites](#)



Secret Decoder Ring Dude

- ▶ Another spam that looks easy

Online Pharmacy - 24/7 Customer Care

Hundreds of products for dozens of ailments. We carry everything from Pain Relief to Skin Care products. Our most popular include:

- Viagra - Proven sexual aid to enhance performance
- Soma - The best in muscle relaxation available
- Phentermine - Safe, proven way to reduce weight
- ... and more!

[Visit the Online Pharmacy for your medical needs](#)

- ▶ Is it?

Please unsubscribe me

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.

Disadvantages:

- 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.

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.

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.

SYSTEM REQUIREMENTS:

Hardware requirements:

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

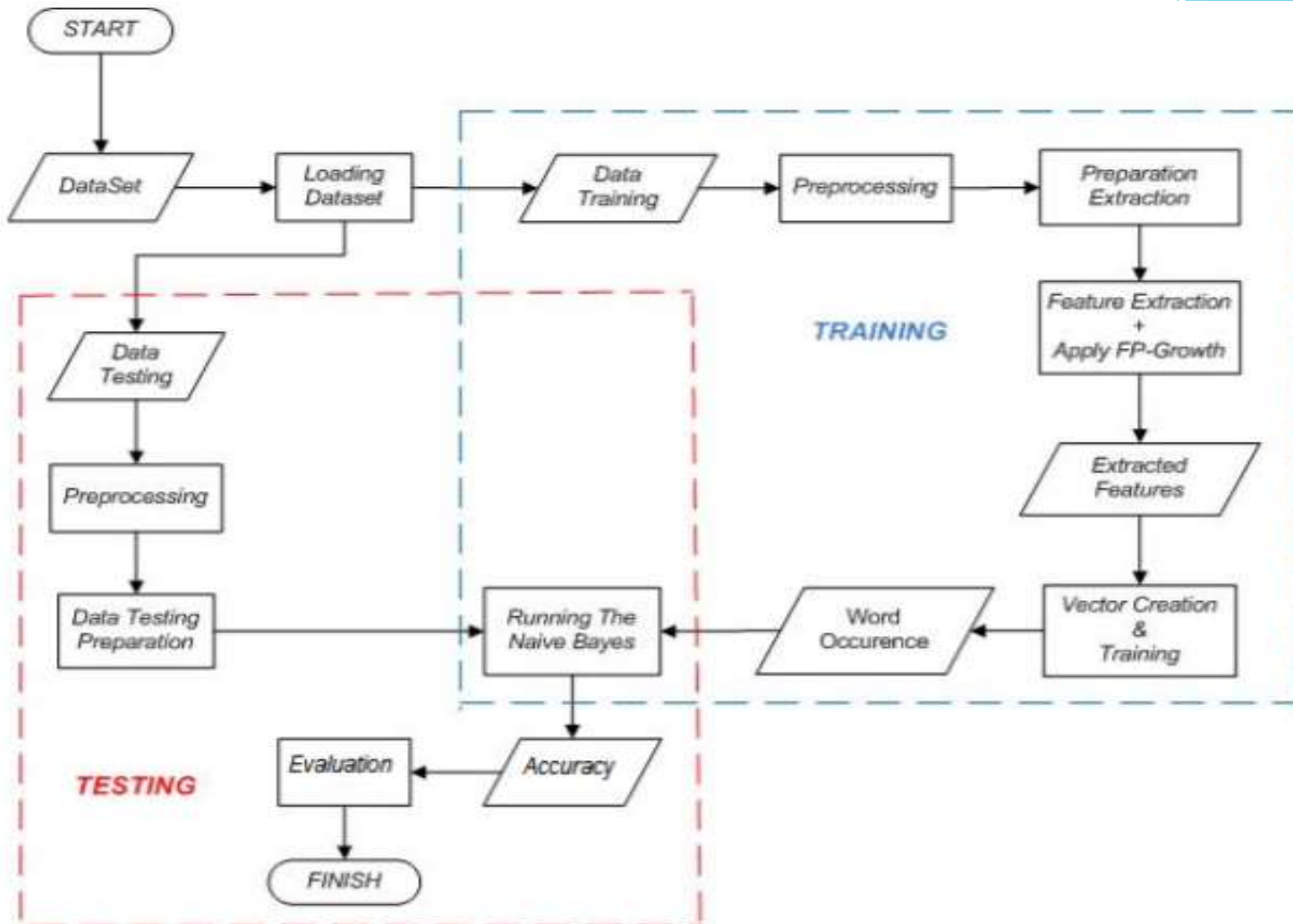
Software requirements:

•Operating system	:	Windows 7/8/10
•Coding Language	:	Python
•IDE	:	Anaconda(spyder, Jupiter), Python IDLE 2.7/3.6
•Database	:	CSV File, TSV File.

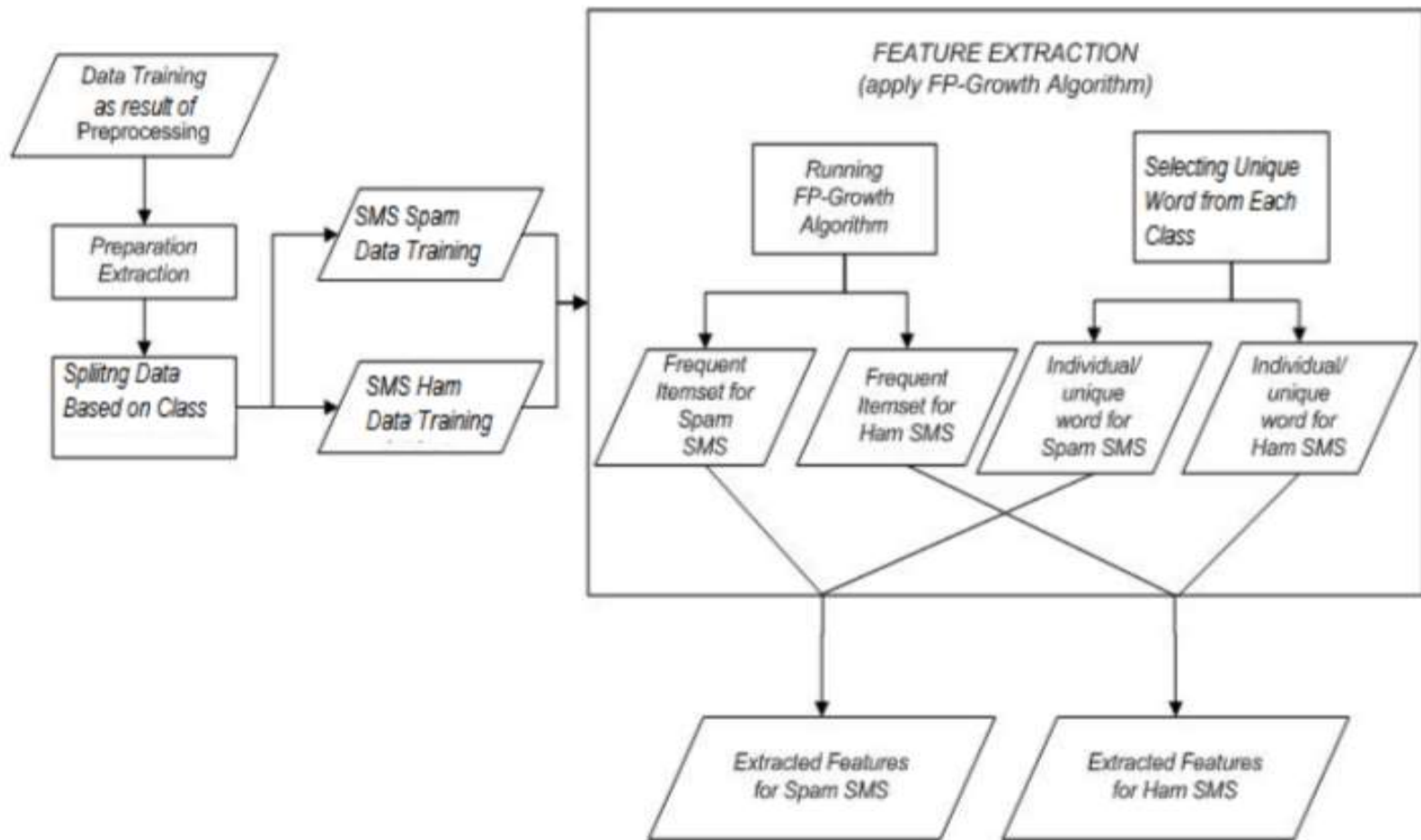
MODULE DESCRIPTION:

- Importing the Libraries
- Load Data sets
- Data Preprocessing
- Feature Extraction (FP-Growth)
- Vector Creation
- Classification
- Naïve Bayes Algorithm
- Finding the accuracy
- Output

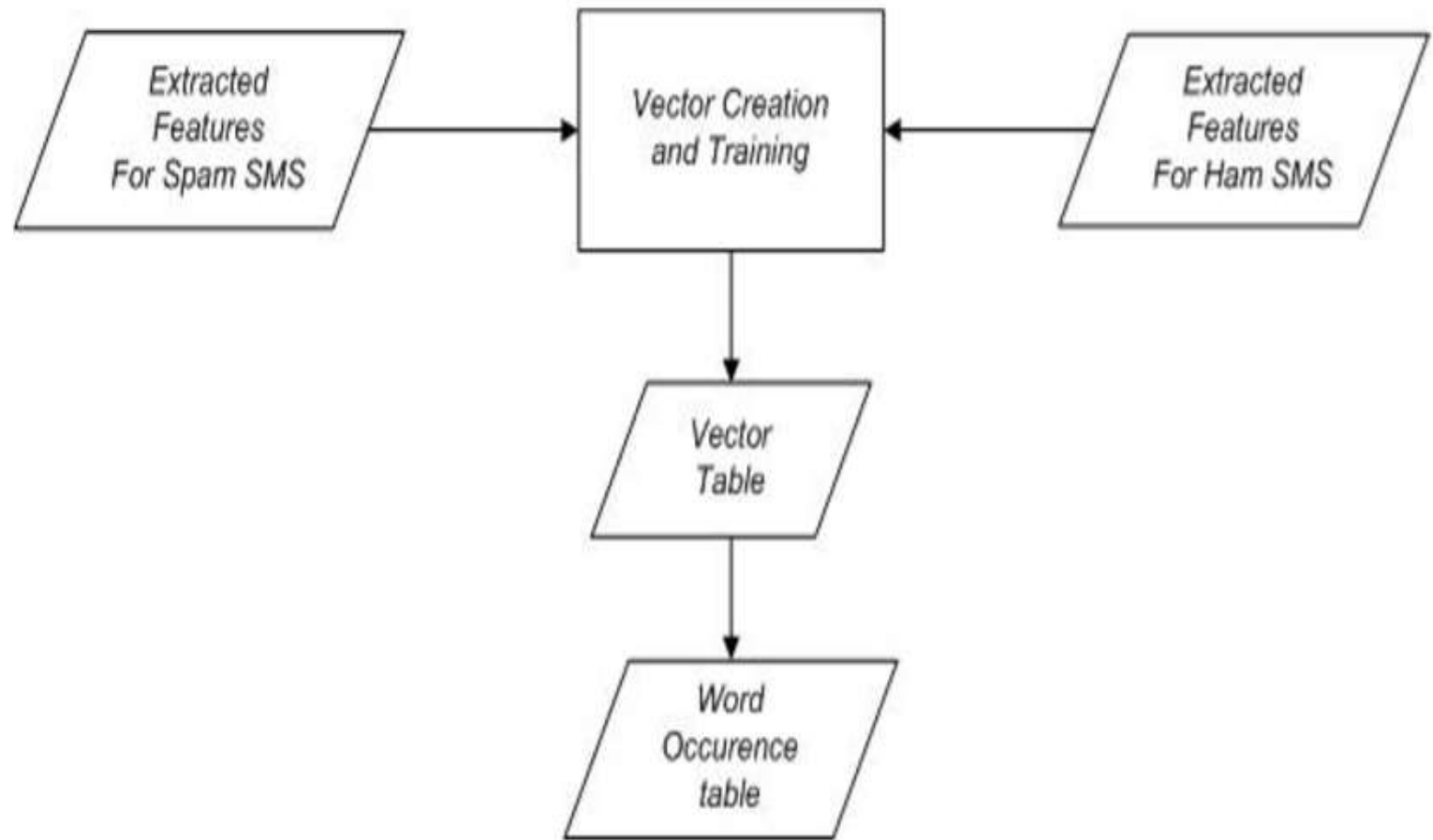
System Architecture



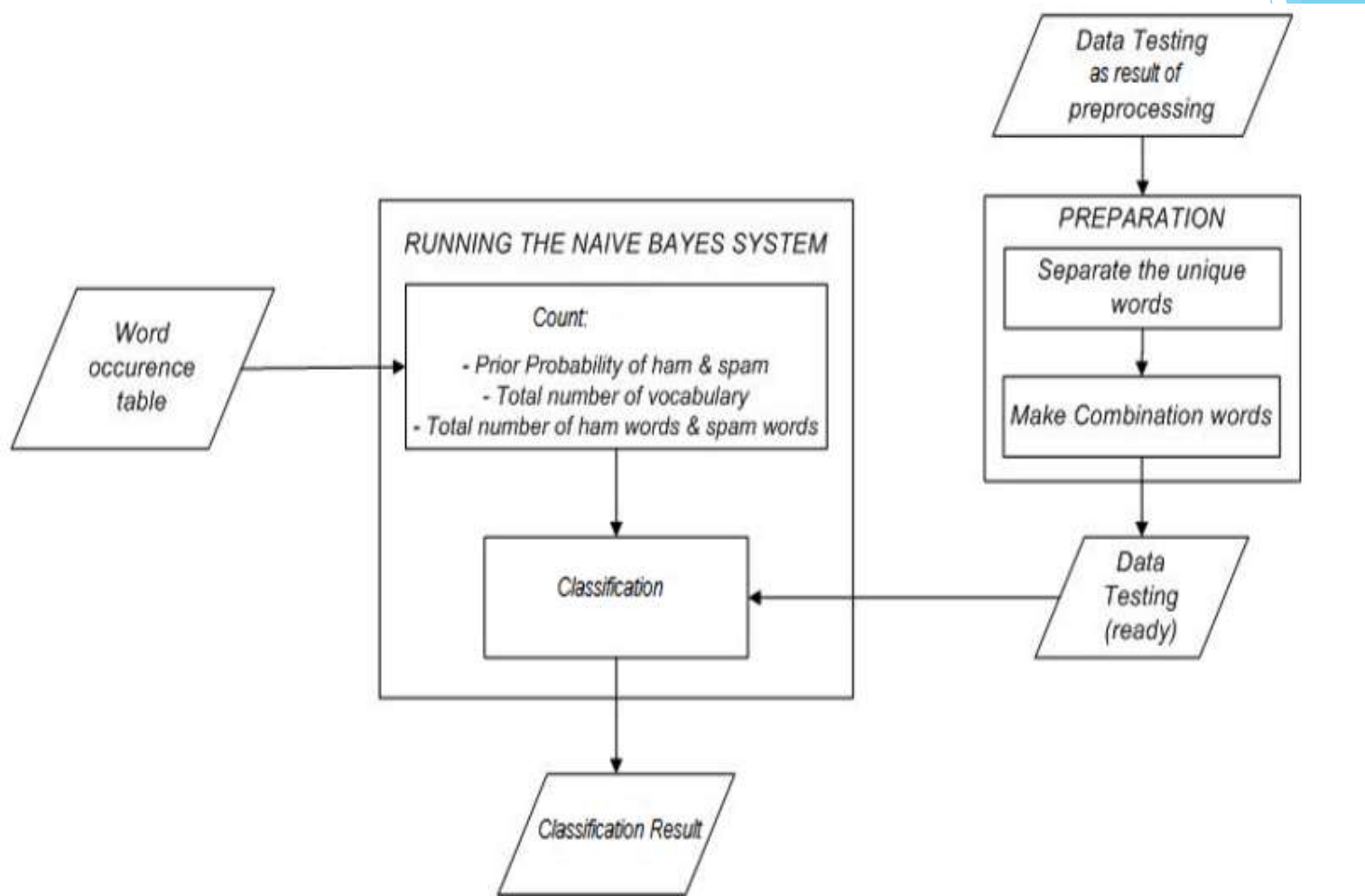
Feature Extraction



Vector Creation and Training



Running the Naive Bayes System



Naïve Bayes Algorithm

- ▶ Want to find $p(\text{spam}/\text{ham})$
- ▶ Use Bayes Rule:
To find the probability

$$1. P(\text{spam}/\text{yes}) = \underline{p(\text{yes}/\text{spam})} * \underline{p(\text{yes})}$$

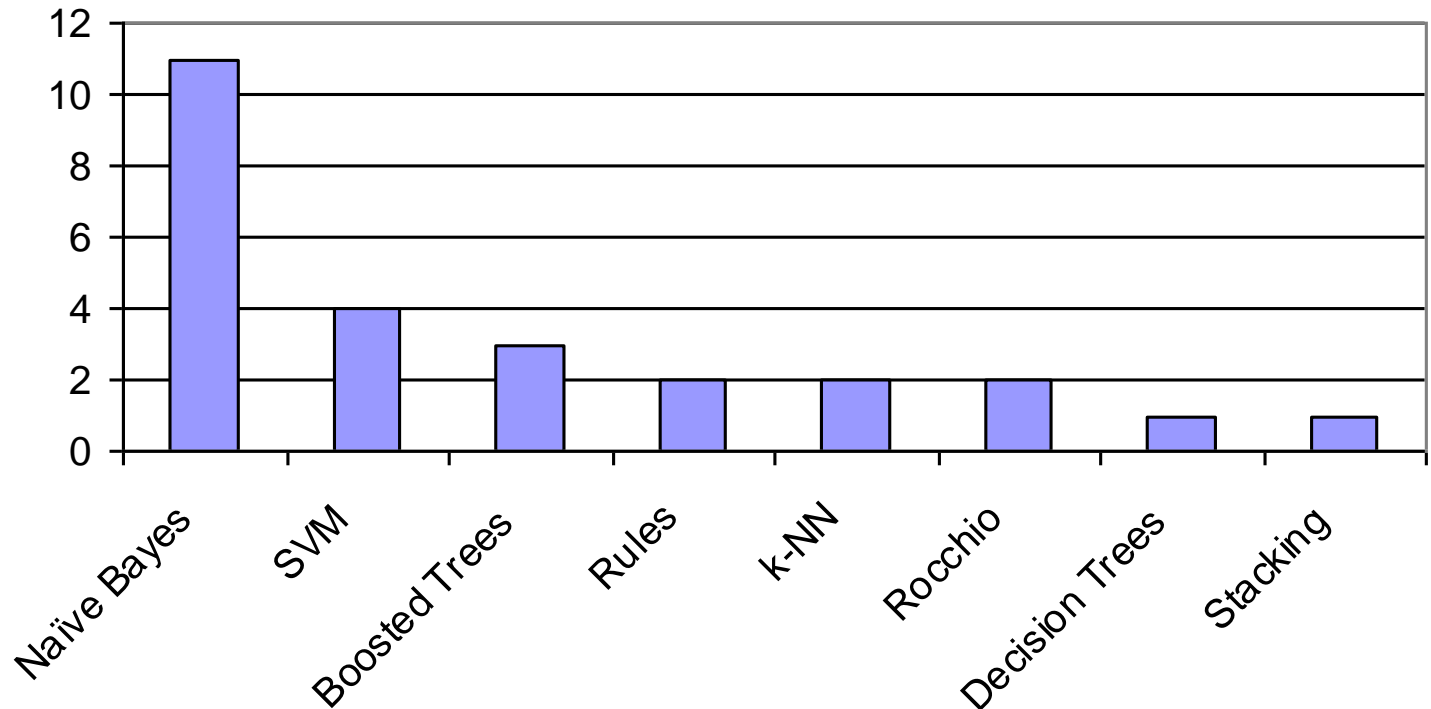
$p(\text{spam})$

$$2. p(\text{spam}/\text{no}) = \underline{p(\text{no}/\text{spam})} * \underline{p(\text{no})}$$

$p(\text{no})$

- ▶ Assume independence: probability of each word independent of others

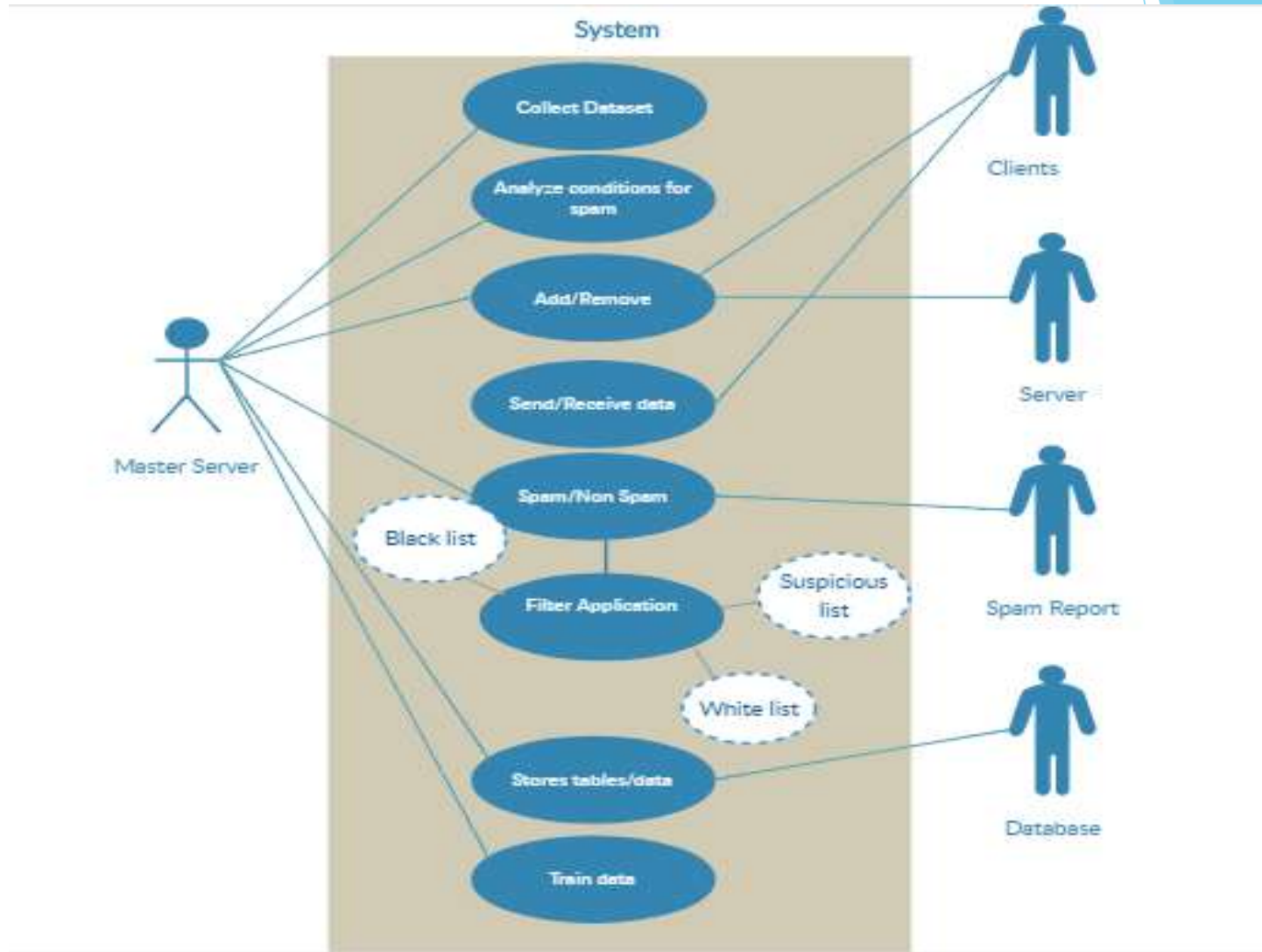
Algorithms Used in Spam Detection



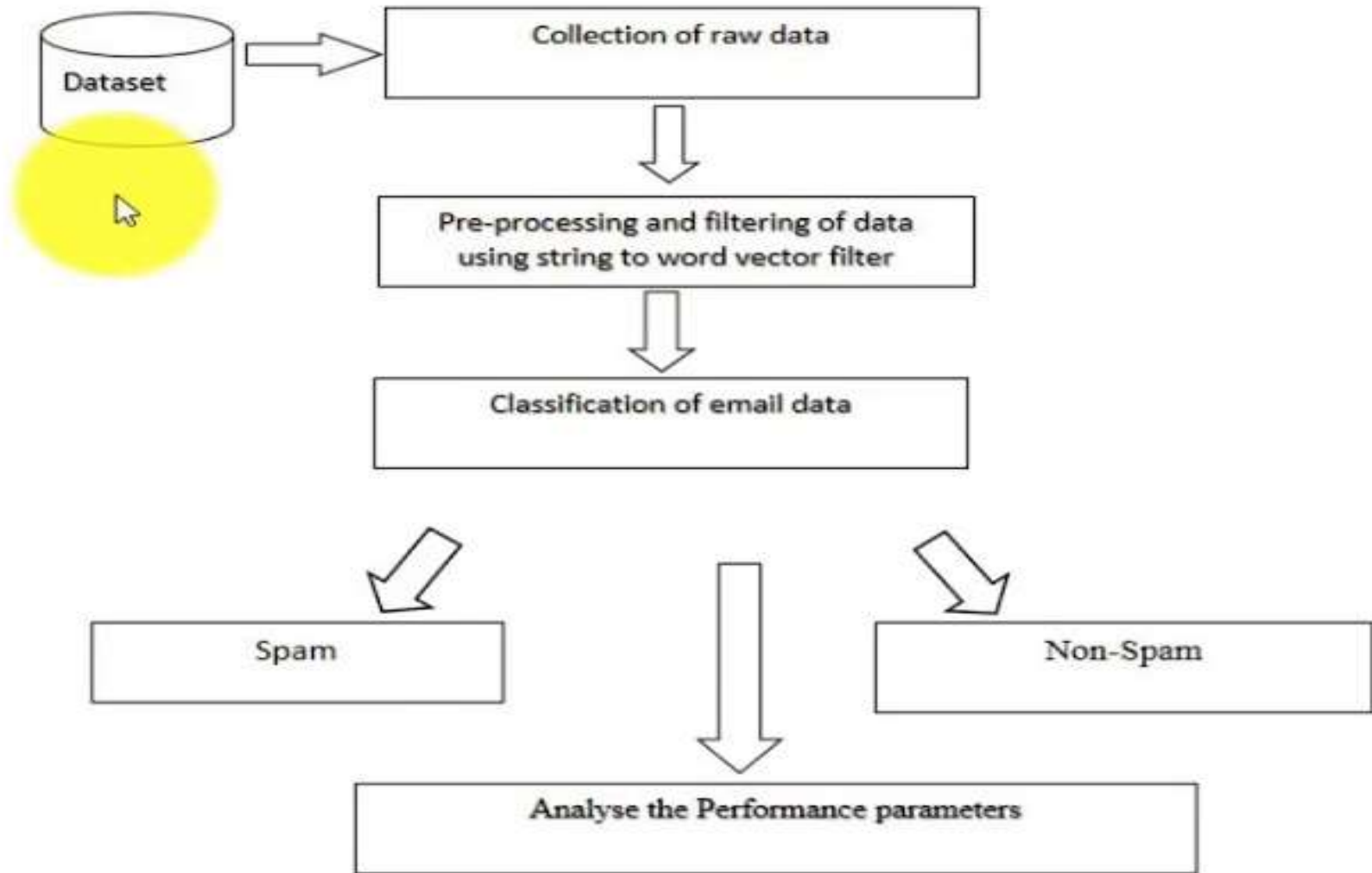
- Naïve Bayes reported to do very well
- More complex algorithms have some gain

UML DIAGRAMS

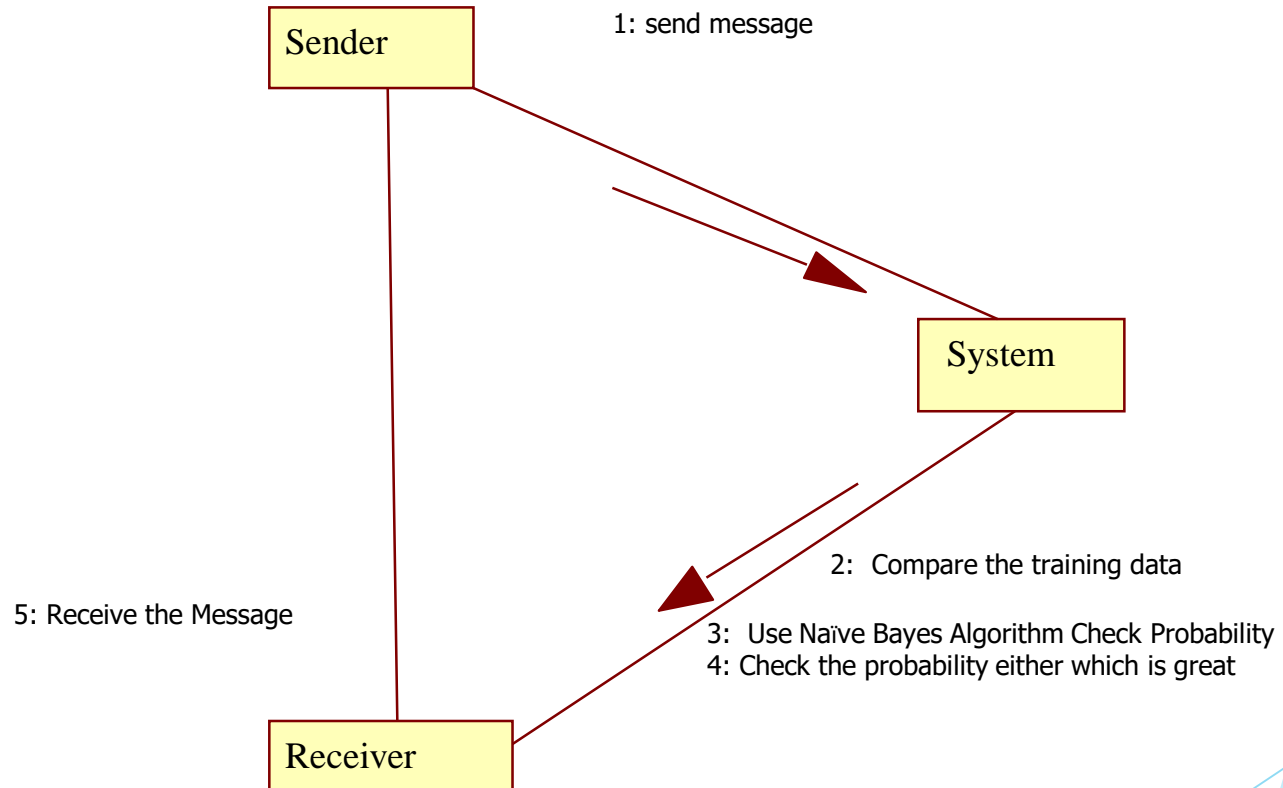
USECASE DIAGRAM



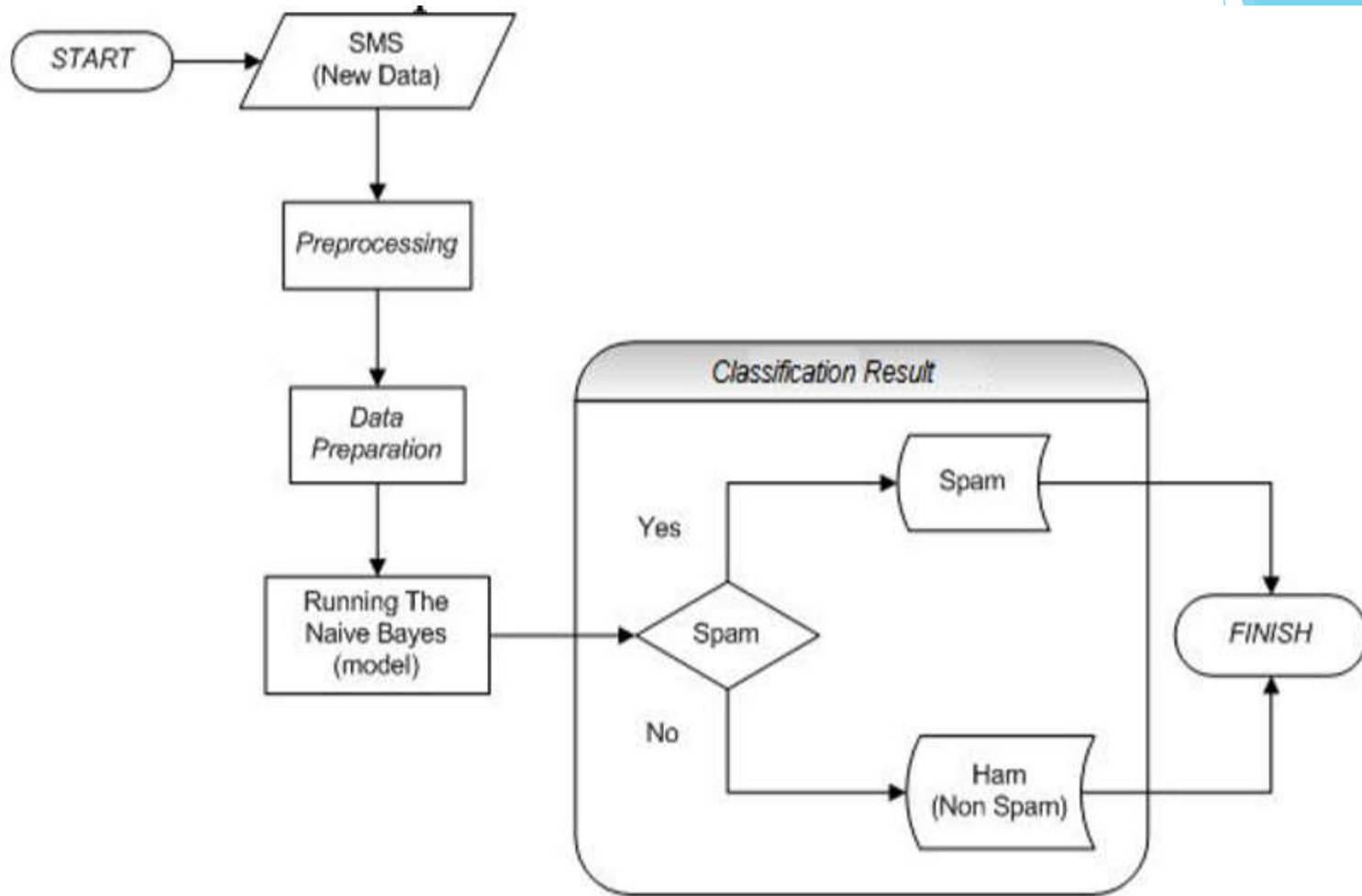
Activity diagram



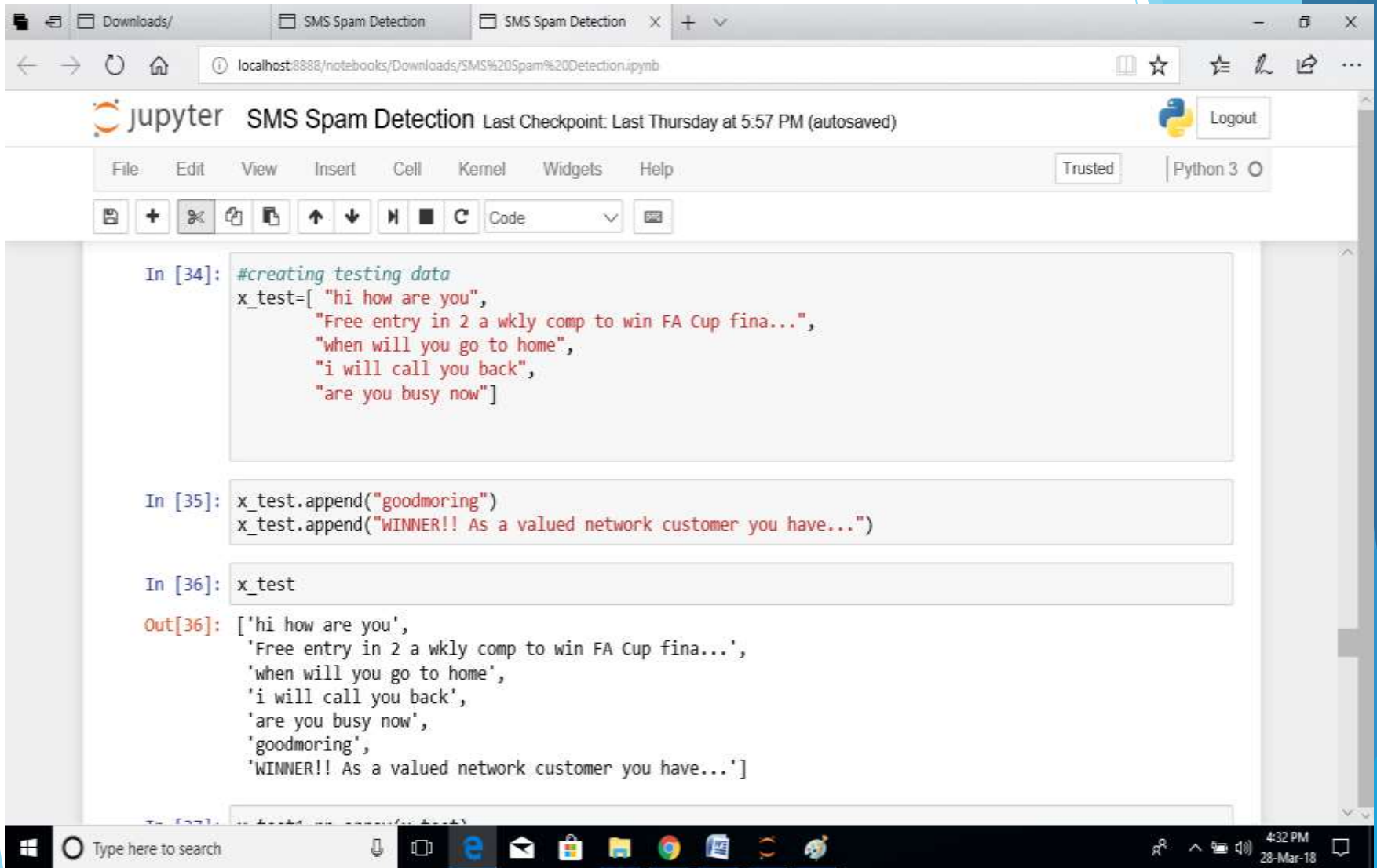
Collaboration Diagram



New SMS prediction process:



New Input Data



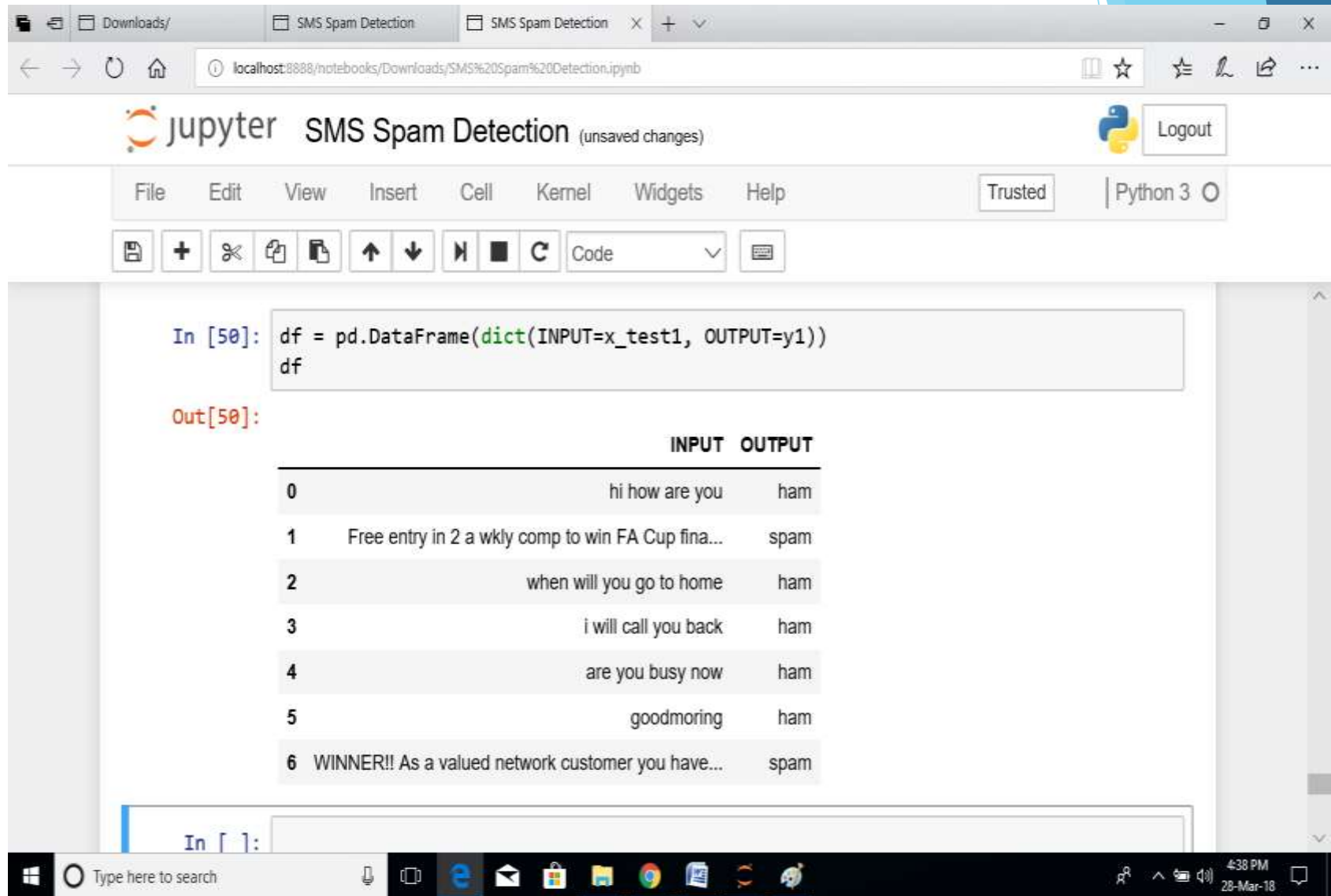
The screenshot shows a Jupyter Notebook titled "SMS Spam Detection" running on a local host. The notebook has a menu bar with File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. Below the menu bar is a toolbar with icons for saving, adding cells, and other functions. The notebook content consists of three input cells and one output cell. The first input cell (In [34]) creates a list named x_test with five string elements. The second input cell (In [35]) appends two more strings to the list. The third input cell (In [36]) prints the list. The output cell (Out[36]) shows the resulting list with seven elements.

```
In [34]: #creating testing data
x_test=[ "hi how are you",
          "Free entry in 2 a wkly comp to win FA Cup fina...",
          "when will you go to home",
          "i will call you back",
          "are you busy now"]

In [35]: x_test.append("goodmoring")
x_test.append("WINNER!! As a valued network customer you have...")

In [36]: x_test
Out[36]: ['hi how are you',
          'Free entry in 2 a wkly comp to win FA Cup fina...',
          'when will you go to home',
          'i will call you back',
          'are you busy now',
          'goodmoring',
          'WINNER!! As a valued network customer you have...']
```


Output of our project



The screenshot displays a Jupyter Notebook window titled "SMS Spam Detection" with "(unsaved changes)". The browser address bar shows "localhost:8888/notebooks/Downloads/SMS%20Spam%20Detection.ipynb". The notebook interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for saving, adding cells, and running code. The current cell is a code cell with the following input:

```
In [50]: df = pd.DataFrame(dict(INPUT=x_test1, OUTPUT=y1))
df
```

The output of this cell is displayed below the code:

```
Out[50]:
```

	INPUT	OUTPUT
0	hi how are you	ham
1	Free entry in 2 a wkly comp to win FA Cup fina...	spam
2	when will you go to home	ham
3	i will call you back	ham
4	are you busy now	ham
5	goodmoring	ham
6	WINNER!! As a valued network customer you have...	spam

The bottom of the image shows the Windows taskbar with the search bar and various application icons. The system clock indicates 4:38 PM on 28-Mar-18.

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.

Thank you

