Spam Detection Using Machine Learning

Report Submitted To

Techno India University, West Bengal

For The Partial Fulfilment

Of

Bachelor of Technology (B.Tech)

Degree In

Computer Science & Engineering

By

Name	ID
Ankit Kumar Singh	181001001075
Saurav Tripathi	181001001061
Sanchari Paul	181001001081
Subhajit Gorai	181001001150
Shahil Choudhary	181001001074

TECHNO INDIA
UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
TECHNO INDIA UNIVERSITY, WEST BENGAL,
SALT LAKE, KOLKATA – 700091, INDIA

June 2022

CERTIFICATE

This is to certify that the Dissertation Report entitled, "**Spam Detection Using Machine Learning**" submitted by "**Ankit Kumar Singh, Saurav Tripathi, Sanchari Paul, Subhajit Gorai & Shahil Choudhary**" to Techno India University, Kolkata, India, is a record of bonafide Project work carried out by them under my supervision and guidance and is worthy of consideration for the award of the degree of Bachelor of Technology (B.Tech) in Computer science & Engineering.

Approved By:

_____

Supervisor(s)

Date :

_____

HOD, CSE, Techno India University

Date :

ACKNOWLEDGEMENT

We would first like to thank our thesis supervisor Sucharita Das, Professor, CSE Dept, TIU. She helped us whenever we ran into a trouble spot or had a question about our research or writing.

We take this opportunity to express gratitude to all of the Department faculty members for their help and support.

We also thank our parents for the unceasing encouragement, support and attention and also sense of gratitude to one and all, who directly or indirectly, have bestowed their hand in this thesis.

Ankit Kumar Singh

Saurav Tripathi

Sanchari Paul

Subhajit Gorai

Shahil Choudhary

TABLE OF CONTENTS

ABSTRACT

Spam emails are known as unrequested commercialised emails or deceptive emails sent to a specific person or a company. Spams can be detected through natural language processing and machine learning methodologies. Machine learning methods are commonly used in spam filtering. These methods are used to render spam classifying emails to either ham (valid messages) or spam (unwanted messages) with the use of Machine Learning classifiers. The proposed work showcases differentiating features of the content of documents. There has been a lot of work that has been performed in the area of spam filtering which is limited to some domains. Research on spam email detection either focuses on natural language processing methodologies on single machine learning algorithms or one natural language processing technique on multiple machine learning algorithms . In this Project, a modelling pipeline is developed to review the machine learning methodologies.

Keywords :

- Machine Learning
- Neural Networks
- Naive Bayes
- Logistic Regression

1. INTRODUCTION

Spam E-mails can be not only annoying but also dangerous to consumers.

Spam E-mails can be defined as:

1. Anonymity 2. Mass Mailings 3. Unsolicited

Spam Emails are messages randomly sent to multiple addressees by all sorts of groups, but mostly lazy advertisers and criminals who wish to lead you to phishing sites.

1.1 Objectives

The objectives of identification of Spam Emails are :

- To give knowledge to the user about the fake Emails and relevant Emails.
- To classify that mail spam or not.

1.2 Problem specification

- Unwanted Emails irritating internet connection.
- Critical Emails messages are missed or delayed.
- Millions of compromised computers.
- Billions of dollars lost worldwide
- Identity theft.

- Spam can crash mail servers and fill up hard drives.

1.3 Methodologies

The era of creation of this product includes models i.e., object-oriented model, Prototype Model,waterfall model etc. for making the correct system. water model, the oldest model of creation of the correct system. The product model used by our framework is the cascade model. Cascade model could be a precise and successive way to contend with the merchandise improvement. This incorporates a framework coming up with and displaying that sets up requirements for all the framework parts and distribution of some set of those conditions to programming. Framework building and examination incorporate requirement gathering at the framework level with a modest amount of top-ranking arrangement. Examination info building consolidation would like assortment at the key business level and at the business space level.

2. LITERATURE SURVEY

Email :

Electronic mail is a messaging system that electronically transmits messages across computer networks. Anyone is free to use email services through Gmail, Yahoo or people can even register with an Internet Service Provider (ISPs) and be provided with an email account. Only an internet connection is required, otherwise being a free service.

Spam :

Bulk mails that are unnecessary and undesirable can be classified as Spam Mails. These spam emails hold the power to corrupt one's system by filling up inboxes, degrading the speed of their internet connection.

Spam Detection :

Many spam detection techniques are being used now-adays. The methods use filters which can prevent emails from causing any harm to the user. The contributions and their weaknesses have been identified. There are several methods that are accessible to spam, for example location of sender, it’s contents, checking IP address or space names. Spammers use refined variations to avoid spam identification.

Table 1 : Spam Categories

Categories	Descriptions
Health	The spam of fake medications
Promotional products	The spam of fake fashion items like clothes bags and watches
Adult content	The spam of adult content of pornography and prostitution
Finance & marketing	The spam of stock kiting, tax solutions, and loan packages
Phishing	The spam of phishing or fraud

2.1 Existing Systems

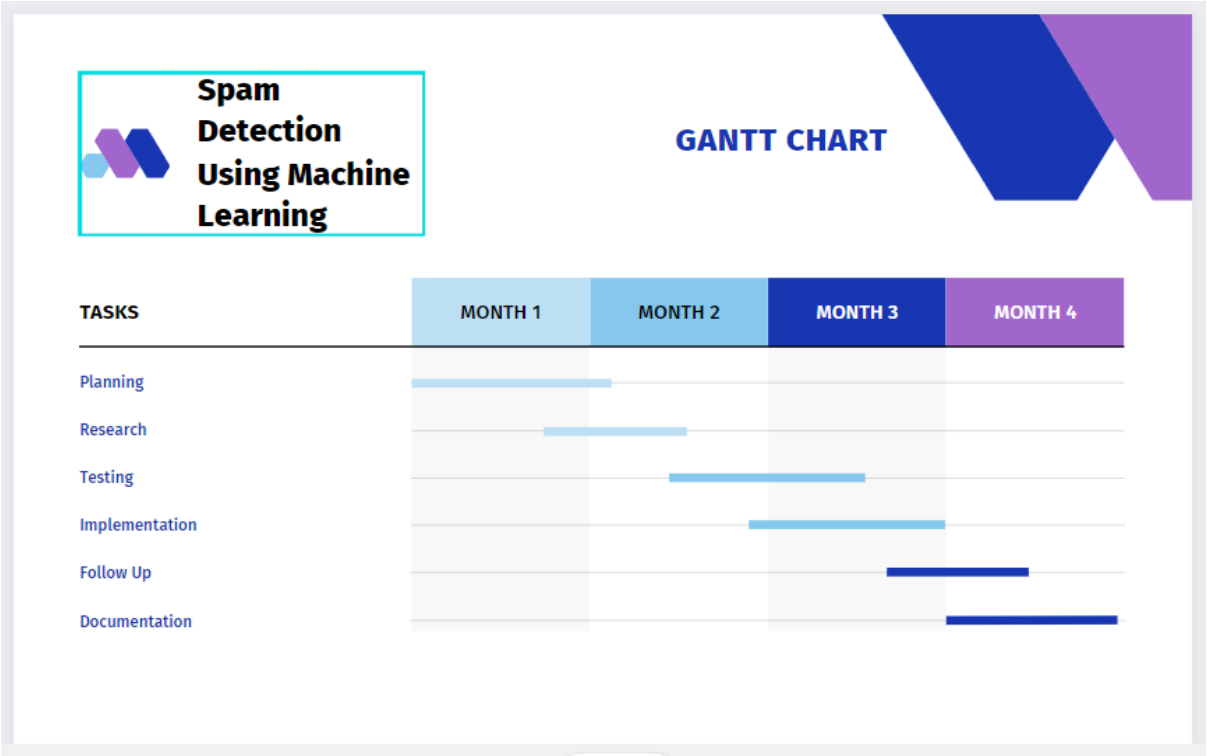
Due to the increase in the number of email users, the amount of spam emails have also risen in number in the past years. Therefore, many researchers have executed comparative studies to see various classification algorithms performances and their results in classifying emails accurately with the help of a number of performance metrics. Hence, it is important to find an algorithm that gives the best possible outcome for any particular metric for correct classification of emails and spam or ham. The present systems of spam detection are reliant on three major methods:-

A. Linguistic Based Methods.

B. Behaviour-Based Methods.

C. Graph-Based Methods

3. PROJECT PLANNING & CHART



4. PROJECT DESCRIPTION

4.1 Software Model

The dataset is taken from SpamAssassin 2500 nonspam messages belong to easy_ham and they should be easily differentiated from spam. Instead of using sophisticated and hybrid models, this study relies on relatively simple classification algorithms to solve this problem like Logistic Regression, Naive Bayes, and Support Vector Machine.

The concept of Neural Networks is also used to select the best activation function for spam detection. The dataset is in the form of HTML files which are converted into plaintext during text preprocessing.

This paper has used two feature sets to find the most optimal feature set and respective models.

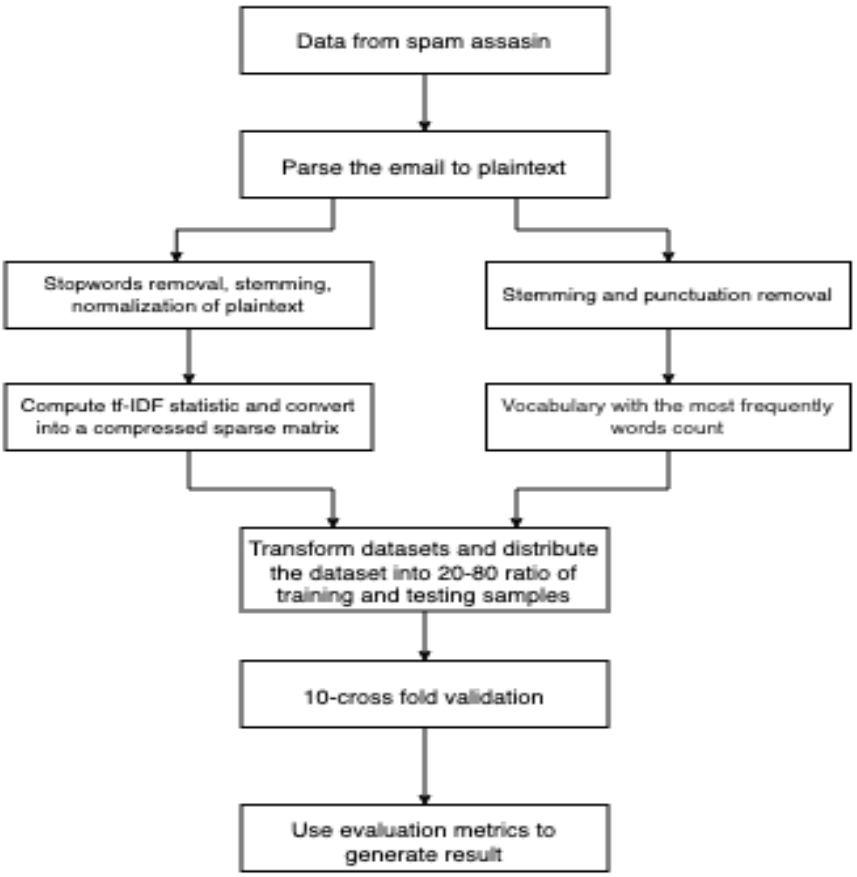
In order to perform efficient operations, Compressed Sparse Row (CSR) is used to feed data to models. Hence, the data is converted into a compressed sparse row matrix format for modelling.

A perfect (or best) model should be the one that reduces underfitting or overfitting. There are three practises for identification.

They are datasets splitting, cross-validation, and bootstrap.

In proposed work to prevent underfitting and overfitting, the modelling results will be evaluated first through a 10-fold cross-validation score, and then evaluated by evaluation metrics of classification.

Figure - Flow Chart of Method



4.2 Software Requirements Specifications (SRS)

Functional Requirement

- 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 spam or not.

Non - Functional Requirement

- Ensure high availability of Email data here datasets.

- Users should get the result as fast as possible.
- It should be easy to use.
- The user is just required to type the words and click the result to be displayed or the user is just required to enter a pair of reasonable sentences.

4.3 Functional Specification

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 it in the form of a matrix.

4.3.1 Functions Performed

```
[1]: import pandas as pd
```

```
[2]: data=pd.read_csv("spam.csv", encoding="latin-1")
```

```
[3]: data.head(5)
```

[3]:

	class	message	sender	receiver
0	ham	Go until jurong point, crazy.. Available only ...	abc@gmail.com	xyz@gmail.com
1	ham	Ok lar... Joking wif u oni...	def@gmail.com	uvw@gmail.com
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	ghi@gmail.com	rty@gmail.com
3	ham	U dun say so early hor... U c already then say...	jkl@gmail.com	iop@gmail.com
4	ham	Nah I don't think he goes to usf, he lives aro...	mno@gmail.com	pok@gmail.com

```
[4]: data.columns
```

```
[4]: Index(['class', 'message', 'sender', 'receiver'], dtype='object')
```

```
[5]: data['class']=data['class'].map({'ham':0, 'spam':1})
```

```
[6]: data.head()
```

[6]:

	class	message	sender	receiver
0	0	Go until jurong point, crazy.. Available only ...	abc@gmail.com	xyz@gmail.com
1	0	Ok lar... Joking wif u oni...	def@gmail.com	uvw@gmail.com
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	ghi@gmail.com	rty@gmail.com
3	0	U dun say so early hor... U c already then say...	jkl@gmail.com	iop@gmail.com
4	0	Nah I don't think he goes to usf, he lives aro...	mno@gmail.com	pok@gmail.com

[7]:
```
#NLP TECHNIQUES
```

[8]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

[9]:
```
cv=CountVectorizer()
```

[10]:
```
X=data['message']
y=data['class']
```

[11]:
```
X.shape
```

[11]:
```
(5244,)
```

[12]:
```
y.shape
```

[12]:
```
(5244,)
```

[13]:
```
X=cv.fit_transform(X)
```

[14]:
```
X
```

[14]:
```
<5244x8426 sparse matrix of type '<class 'numpy.int64'>'
        with 69778 stored elements in Compressed Sparse Row format>
```

[15]:
```
from sklearn.model_selection import train_test_split
```

[16]:
```
x_train, x_test,y_train, y_test=train_test_split(X,y)
```

[17]:
```
x_train.shape
```

[17]:
```
(3933, 8426)
```

[18]:
```
from sklearn.naive_bayes import MultinomialNB
```

[19]:
```
model=MultinomialNB()
```

[20]:
```
model.fit(x_train, y_train)
```

[20]:
```
MultinomialNB()
```

[21]:
```
result=model.score(x_test, y_test)
```

[22]:
```
result=result*100
```

[23]:
```
result
```

[23]:
```
97.86422578184592
```

[24]:
```
import pickle
```

[25]:
```
pickle.dump(model, open('spam.pkl','wb'))
```

[26]:
```
pickle.dump(cv, open("vectorizer.pkl","wb"))
```

[27]:
```
clf=pickle.load(open("spam.pkl","rb"))
```

[28]:
```
clf
```

[28]:
```
MultinomialNB()
```

[29]:
```
#EXAMPLE TO CHECK SPAM DETECTION
```

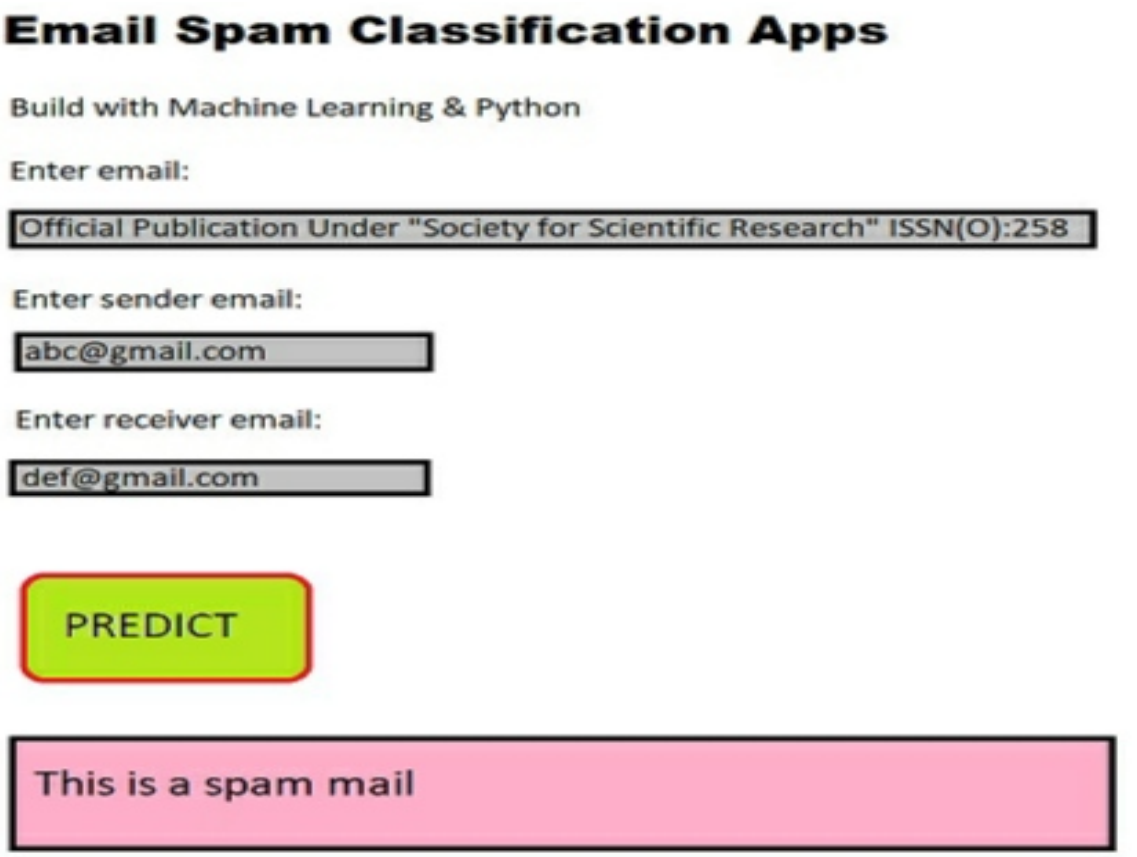
[30]:
```
msg="You Won 500$"
data = [msg]
vect = cv.transform(data).toarray()
my_prediction = model.predict(vect)
print(my_prediction)
```

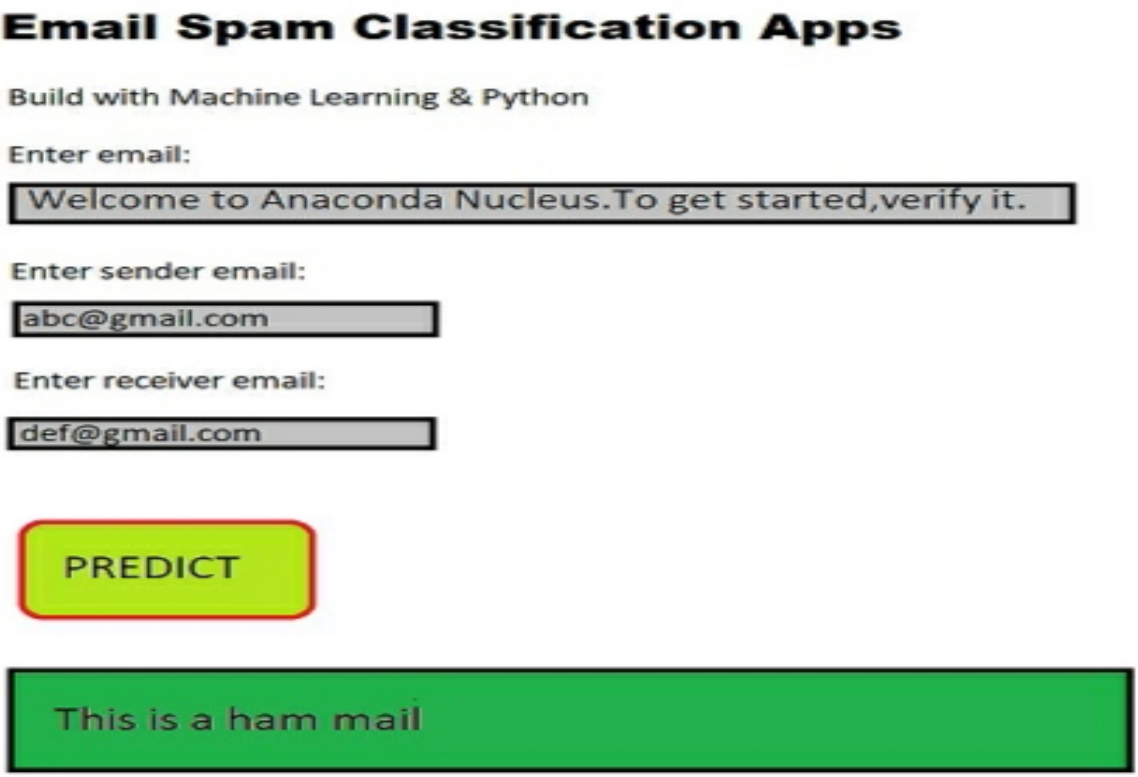
[2]:
```
import pickle
import streamlit as st


model = pickle.load(open('spam.pkl','rb'))
cv=pickle.load(open('vectorizer.pkl','rb'))


def main():
    st.title("Email Spam Classification Apps")
    st.subheader("Build with Machine Learning & Python")
    msg=st.text_input("Enter email:")
    sender=st.text_input("Enter sender email:")
    receiver=st.text_input("Enter receiver email:")
    if st.button("Predict"):
        data=[msg]
        vec=cv.transform(data).toarray()
        prediction=model.predict(vec)
        result=prediction[0]
        if result==1:
            st.error("This is A Spam Email")
        else:
            st.error("This is A Ham Email")

main()
```

4.3.2 Limitations and Restrictions

Our project, therefore spam filter, is capable of filtering mails according to the domain names listed in black list only. Therefore it, at this stage, is not able to filter the spams on the basis of its contents or some other criteria.

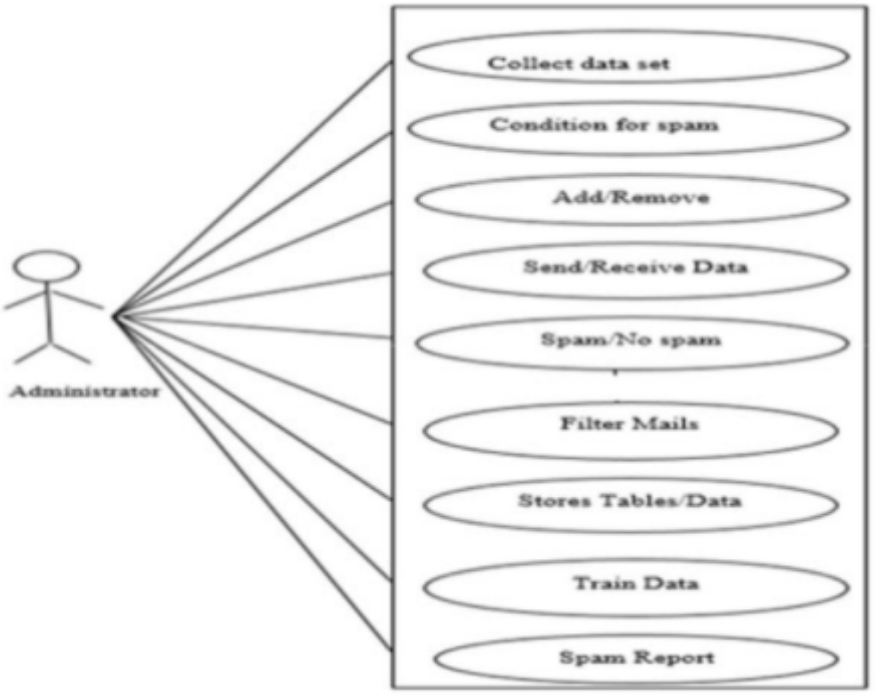
4.3.3 User Interface Design



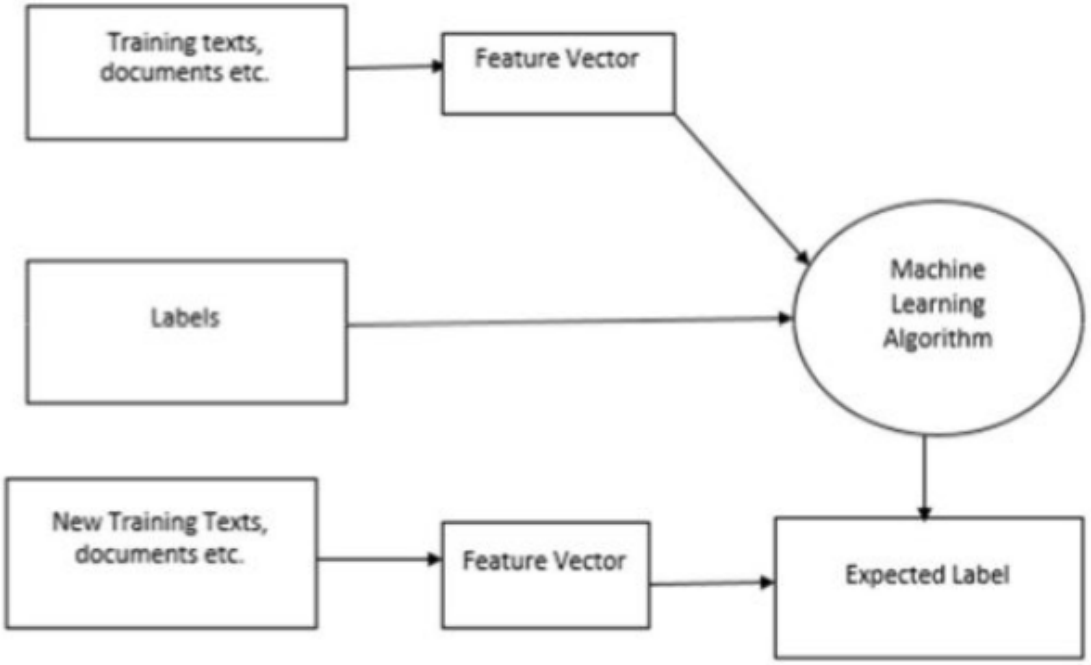


4.4 Design Specification

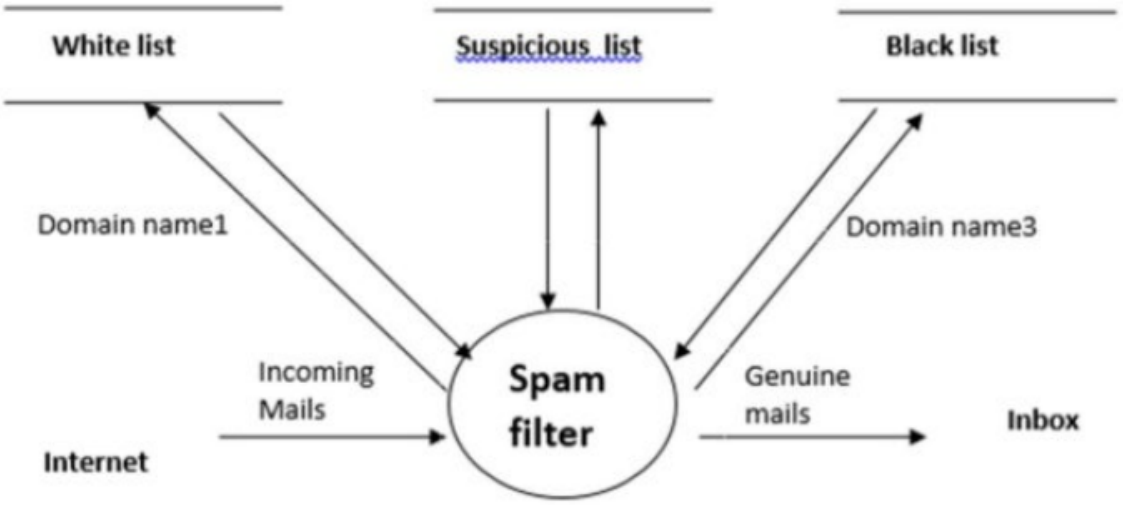
4.4.1 E-R diagram / Use-case diagram (UML)



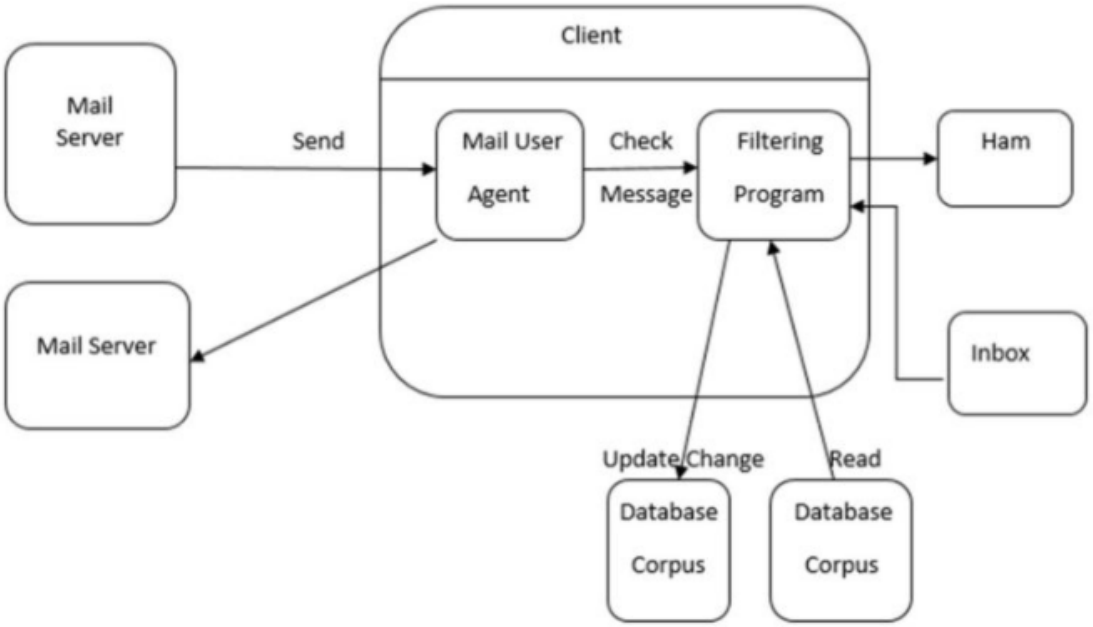
4.4.2 Data Flow Diagrams (DFD)



Level - 1



Level - 2



Level - 3

4.4.3 Data Dictionary

Dataset is a collection of data or related information that is composed of separate elements. A collection of dataset for e-mail spam contains spam and non-spam messages. In this research, two datasets are being used to evaluate the performance of Naive Bayes algorithm to filter email spam.

4.5 Testing

In order to create an algorithm for this, we need to teach our program what a spam email looks like and what non-spam emails look like. . We also need a way to test the accuracy of our spam filter. One idea would be to test it on the same data that we used for training. However, this can lead to a major problem in ML called overfitting which means that our model is too biassed towards the training data and may not work as well on elements outside of that training set. One common way to avoid this is to split our labelled data for training or testing. This ensures we test on different data than we trained on. It's important to note that we need a mix of both spam and non-spam data in our data sets, not just the spam ones. We really want our training data to be as similar to real email data as possible.

5. IMPLEMENTATION ISSUES

A subtle issue with Naive-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.

6. EXPERIMENTAL RESULTS

The evaluation criteria is simply based on the following evaluation metrics :

- Accuracy
- Precision

• Recall
• F1 score

These four factors comprehend the performance of a model with the feature set. In the figure above, it is shown how different models perform with these respective metrics.

As shown by the accuracy graphs it can be seen that the artificial neural network has the highest detection rate of whether a file is spam or ham. Also as shown by recall and F-Score it can be seen that the Neural Network out performs every other model.

However results can also be
seen in terms of precision logistic regression is the better, however it's not the best model as its poor performance compared to others.

Table 1 and Table 2 showcases the output of the results of feature set 1 and feature set 2 with the models respectively. cv_score_mean refers to cross validation
score, and is used to verify accuracy results. cv_score_std refers to deviation in cross validation and also how much overfitting is there in the model.

Among all models using the feature 1 stopwords + n-gram + tf-idf as shown in Table 2, Neural Network using tanh activation function achieved maximum accuracy and viz. 98.69%. Logistic Regression got the highest precision 99.33% so false-positives are least there.

Among all models using the feature 2 word-count as

shown in Table 3, Neural Network using tanh activation function achieved maximum accuracy and viz. 99.02%. Logistic Regression got the highest precision 99.33% so false-positives are least there but it's score and recall are less than a Neural Network.

7. CONCLUSION

As shown in Figure , all the models based on the feature set 2 most-frequent-word-count have higher accuracy and F1 score than those based on the feature set 1 stop words + n-gram + tf-IDF.

If the use case is to introduce a beta version of an email spam detector like no-spam in the inbox. In this case, the model: Neural Network with tanh activation function and the feature set 1 stop words + n-gram + tf-IDF serves this purpose. According to the graphs in Figure , if the use case is to introduce an email spam detector to reduce bad user experience in searching for important emails from junk mailboxes and filtering spam from the inbox.

In this case, a Neural Network with a feature set 2 - ‘most frequent word count’ gives a better user experience in general. The future work includes testing the model with various standard datasets. This research proposes that the outcome that is obtained should be compared with additional spam datasets from various sources. Also, more classification and feature algorithms should be analysed with email spam datasets.

8. FUTURE SCOPE

There is a wide scope of enhancement in our project. Following enhancements can be done: Filtering of spams can be done on the basis of its contents. The spam email classification is very important in classifying emails and to separate emails that are spam or non-spam. This method can be used by big organisations to distinguish good mails that are only the mails they wish to receive.

9. REFERENCES

1. M. Awad, M. Foqaha Email spam classification using hybrid approach of RBF neural network and particle swarm optimizationInt. J. Netw. Secur. Appl., 8 (4) (2016).

2. D.M. Fonseca, O.H. Fazzion, E. Cunha, I. Las-Casas, P.D. Guedes, W. Meira, M. Chaves.

3. Measuring characterising, and avoiding spam traffic costs IEEE Int. Comp., 99 (2016).