AI PROJECT REPORT

SPAM EMAIL DETECTION USING PYTHON

SUBMITTED TO :

MR.DIPEN SAINI

SUBMITTED BY :

GOPU SHANMUKHA DATTA ,12104966

S.KRISHNA SAI MANIDEEP ,12105054

M.VENU TEJASWI,12100671

INTRODUCTION :

1.1 BACKGROUND

Today, spam has become a big internet issues. Recent 2017, the statistic shown spam accounted for 55% of all e-mail messages, same as during the previous year. Spam which is also known as unsolicited bulk email has led to the increasing use of email as email provides the perfect ways to send the unwanted advertisement or junk newsgroup posting at no cost for the sender. This chances has been extensively exploited by irresponsible organizations and resulting to clutter the mail boxes of millions of people all around the world.

Evolving from a minor to major concern, given the high offensive content of messages, spam is a waste of time. It also consumed a lot of storage space and communication bandwidth. End user is at risk of deleting legitimate mail by mistake. Moreover, spam also impacted the economical which led some countries to adopt legislation.

In this project, a machine learning technique is used to detect the spam message of a mail. Machine learning is where computers can learn to do something without the need to explicitly program them for the task. It uses data and produce a program to perform a task such as classification. Compared to knowledge engineering, machine learning techniques require messages that have been successfully pre-classified. The pre-classified messages make the training dataset which will be used to fit the learning algorithm to the model in machine learning studio.

1.2 PROBLEM STATEMENT

A tight competition between filtering method and spammers is going on per day, as spammers began to use tricky methods to overcome the spam filters like using random sender addresses or append random characters at the beginning or end of mails subject line. There is a lack of machine learning focuses on the model development that can predict the activity. Spam is a waste of time to the user since they have to sort the unwanted junk mail and it consumed storage space and communication bandwidth. Rules in other existing must be constantly updated and maintained make it more burden to some user and it is hard to manually compare the accuracy of classified data.

1.3 OBJECTIVES

There are four objectives that need to be achieved in this project:

i. To study on how to use machine learning techniques for spam detection.

ii. To modify machine learning algorithm in computer system settings.

iii. To leverage modified machine learning algorithm in knowledge analysis software. iv. To test the machine learning algorithm real data from machine learning data repository.

1.4 PROJECT SCOPE AND LIMITATION OF WORK

1.4.1 PROJECT SCOPE

This project needs a coordinated scope of work. These scopes will help to focus on this project. The scopes are:

i. Modified existing machine learning algorithm.

ii. Make use and classify of a data set including data preparation, classification and visualization.

iii. Score of data to determine the accuracy of spam detection

1.4.2 LIMITATION OF WORK

The limitation of this project are:

i. This project can only detect and calculate the accuracy of spam messages only ii. It focus on filtering, analysing and classifying the messages.
iii. Do not block the messages.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter discusses about the literature review for machine learning classifier that being used in previous researches and projects. It is not about information gathering but it summarizes the prior research that related to this project. It involves the process of searching, reading, analysing, summarising and evaluating the reading materials based on the project.

Literature reviews on machine learning topic have shown that most spam filtering and detection techniques need to be trained and updated from time to time. Rules also need to be set for spam filtering to start working. So eventually it become burdensome to the user.

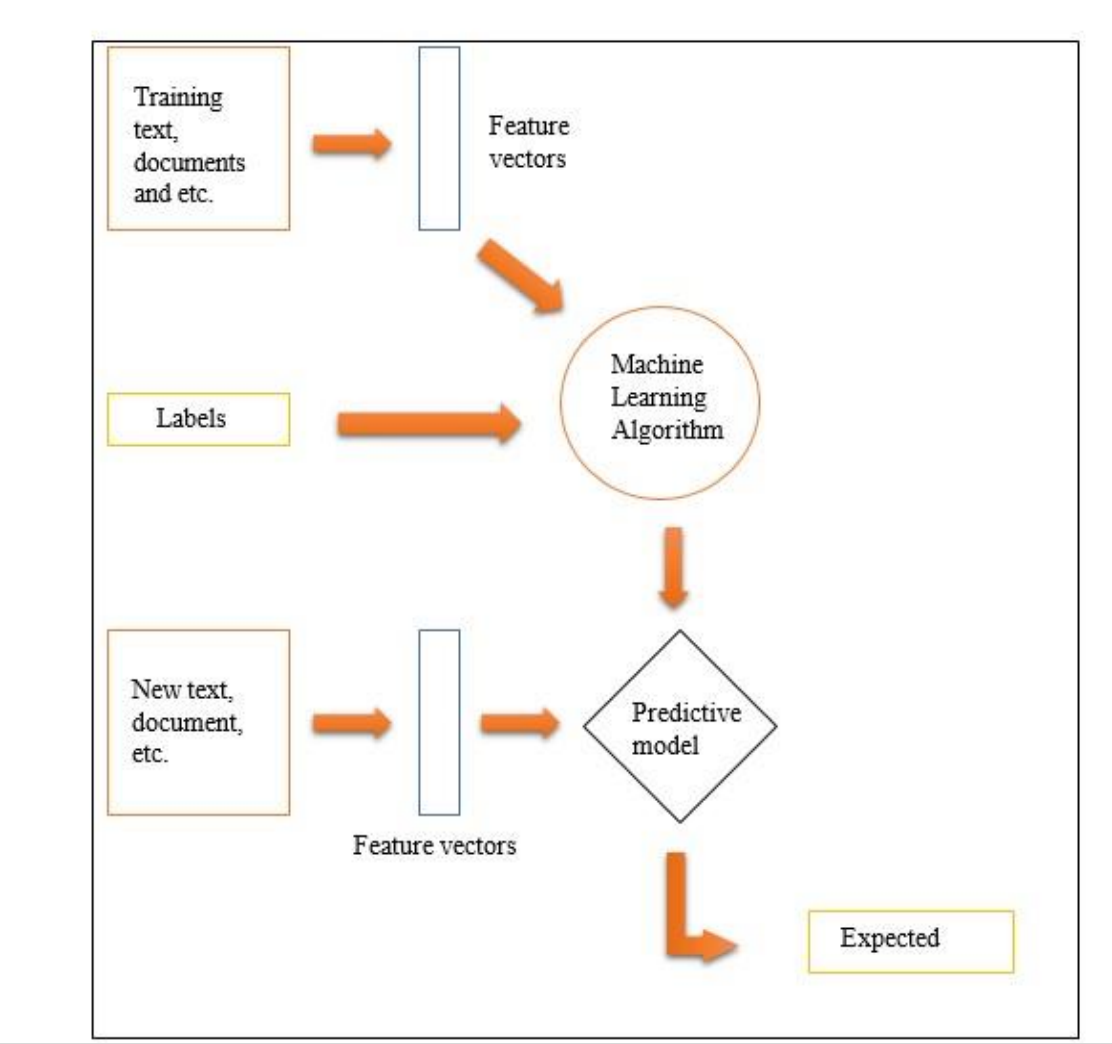
2.2 MACHINE LEARNING

In this project, existing machine learning algorithm is used and modified to fit the need of project. The reasons are because machine learning algorithm is adept at reviewing larges volume of data. It is typically improves over time because of the ever-increasing data that are processed. It gives the algorithm more experience and be used to make better predictions.

Machine learning allows for instantaneous adaption without human intervention. It identifies new threats and trends and implements the appropriate measures. It is also save time as it is it automated nature.

2.3 SPAM DETECTION

In theory, spam detection can be implemented at any location and multiple stages of process can occur at the same time. Figure 1 shows the spam detection process.



2.4RELATED WORK

Most research has been conducted into detecting and filtering spam email using a variety of techniques.

TITLE	TECHNIQUE	REMARK
Spam Filtering Using K-NN	K-nearest Neighbors	Computationally intensive, especially when the size of the training set grows.
A review of machine learning approaches to Spam filtering	Bayesian Filters	Require training period before it starts working well.
SVM-Based Spam Filter with Active and Online Learning	Support Vector Machines	Select the most useful example for labeling and add the labeled example to training set to retrain model.
A survey on spam detection techniques	Artificial Neural Network	Must be trained first to categorize emails into spam or non-spam starting from the particular data sets.

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

This chapter will explain the specific details on the methodology being used in order to develop this project. Methodology is an important role as a guide for this project to make sure it is in the right path and working as well as plan. There is different type of methodology used in order to do spam detection and filtering. So, it is important to choose the right and suitable methodology thus it is necessary to understand the application functionality itself.

3.2 IMPLEMENTATION AND CODING PHASE

This project is developed by using Python Language and combining with the Vowpal Wabbit algorithm. Azure machine learning studio are as the platform to develop the project. It contains important function for preprocessing the dataset. Then, the dataset is going to be used to train and test either the model of the machine learning achieve the objectives of the project.

3.3 PROJECT REQUIREMENT AND SPECIFICATION

System requirement is needed in order to accomplish the project goals and objectives and to assist in development of the project that involves the usage of hardware and software. Each of these requirements is related to each other to make sure that system can be done smoothly.

3.3.1 HARDWARE

The usage of hardware is as below:

No.	Hardware	Type	Description
1.	Laptop	Acer Aspire E 14	• Processor: Intel Core i5, 7th Gen • OS version: Windows 64 bit • RAM: 8 GB
2.	Printer	HP Deskjet 2135	• Printing document
3.	Printed paperwork	A4 paper	• Used to study on how to implement this project from past paperwork

Table 3.1: Hardware used

It is better to use high performance processor to avoid any problem while doing this project. Machine learning project required a high speed processor for a better performance to train a large amount of data.

3.3.2 SOFTWARE

The usage of software in this project is as below:

No.	Software	Description
1.	Microsoft Azure	Machine learning platform Deploy models Run models in cloud
2.	Google Chrome	Used to run web based system
3.	Microsoft Word 2016	Creating and editing report
4.	Microsoft PowerPoint 2016	For presenting finding and result of the project
5.	Github	Get dataset
6.	Kaggle	Get dataset

7.	UCI machine learning repository	Get dataset
8.	Snipping Tool	Captures and screenshot images
9.	WinZip	Extract the data
10.	Visual Studio	Implementation and deployment

Figure 3.6: List of data set by Kaggle

3.4.3 PROCESS MODEL

Process model is a series of steps, concise description and decisions involved in order to complete the project implementation. In order to finish the project within the time given, the flows of project need to be followed. The framework below shows how the overall flow of this project in order to separate between a spam and ham message.

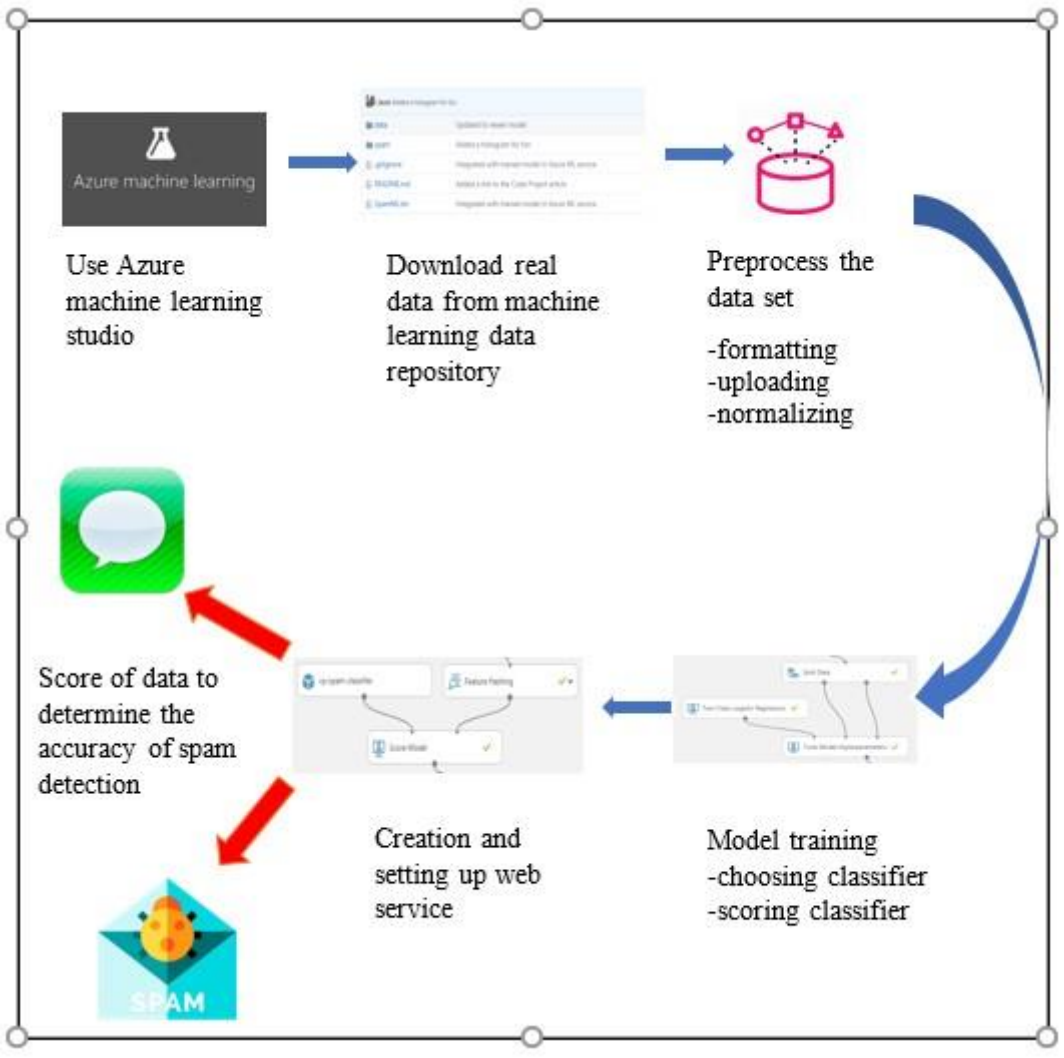


Figure 3.7: Process framework

3.4.4 DATA MODEL

As for data model, it refers to the documenting a complex system and data flow between different data elements and design as an easily understood diagram using text and symbol. The data flow below shows how the data flow of these project in order to detect the spam messages and classify them into two separate type which is spam and ham message.

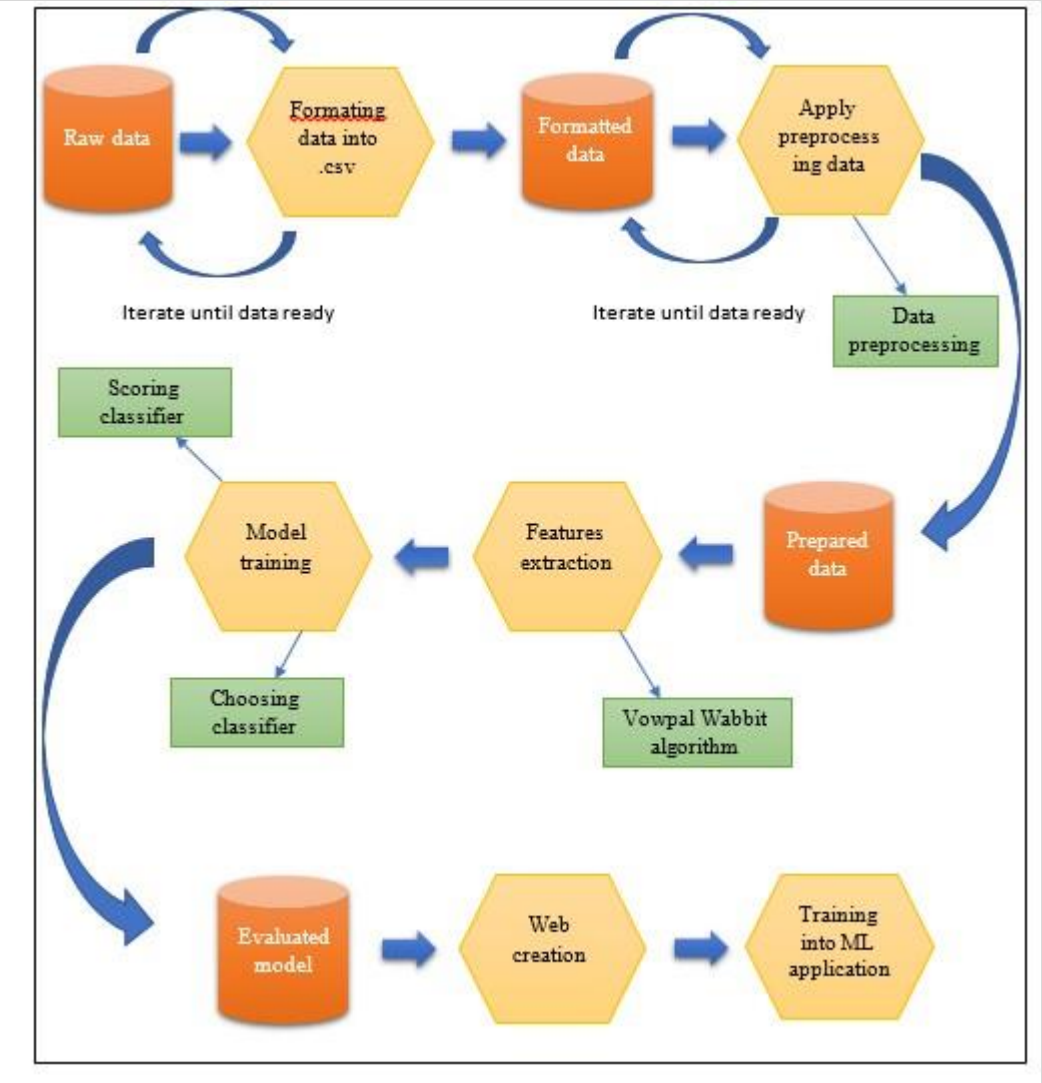


Figure 3.8: Data model flow

4. RESULT

a.CLASSIFICATION AND PROBABILITY

The probability of classification is measured by counting the number of spam flags. According to Dr. Matt Peters, google has examined a great number of potentials factors that predicted that a site might be penalized or banned due to spam [9]. Each flag has its own warning sign that indicates the message as spam. So, to calculate this probability, spam score will records the quantity of flags that triggers the data. Hence, the graph below shows the relationship that numbers of flags effect the probability of classification type. The overall likelihood of spam increases as the number of flags increases.

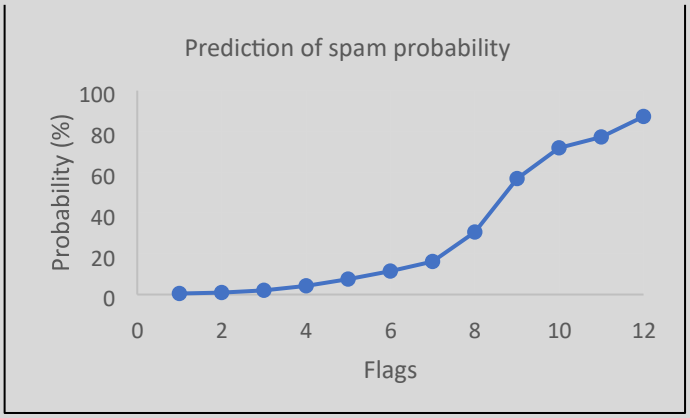


Figure 4.21: Relationship between number of flags and classification type

b. ELAPSED TIME AND MESSAGE COUNT

Elapsed time is time different or amount of time between the beginning and the end of execution process. In simplest terms, elapsed time is the processing time of a process or event. In this project, both elapsed time and message count are taken into consideration in order to score the accuracy. This is to ensure the efficiency of model by decreasing the processing time even when the messages counts are large. As shown in the graph below, it shows the comparison of elapsed time using same messages between four different tools.

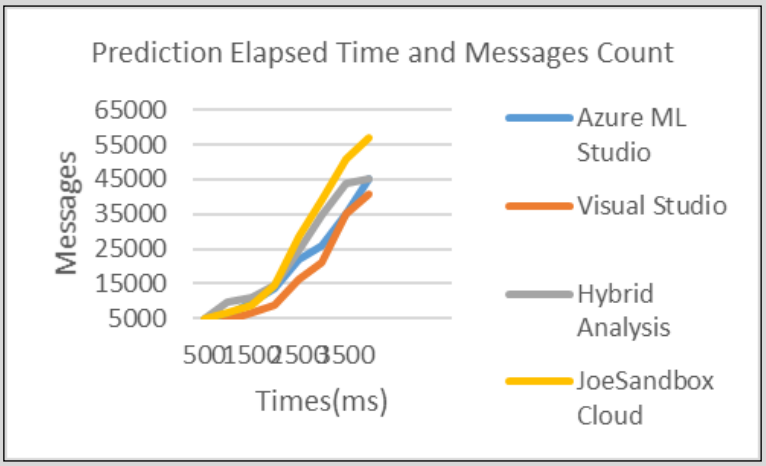


Figure 4.22: The relationship between message count and elapsed time

c. Accuracy and Message Count

Based on research, the message count or frequency of words is calculated in order to get the most accurate percentage of accuracy. This is because, the messages are the important element to test spam detection. Figure below shows that all the tools used verified that accuracy of detection affected by the messages count.

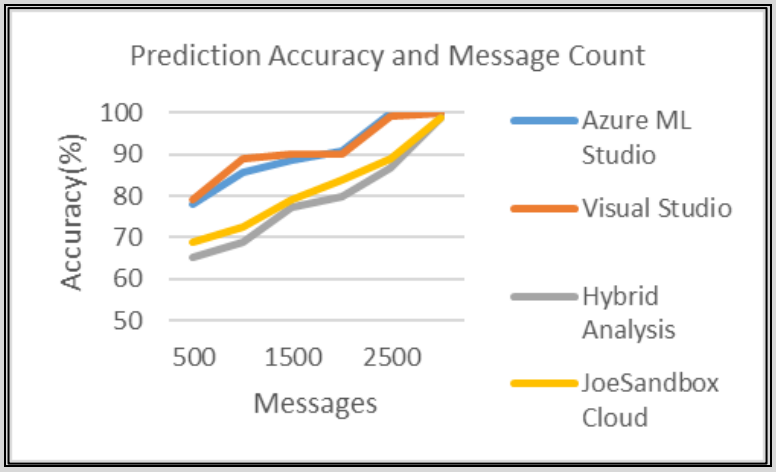


Figure 4.23: Relationship between accuracy and message count

d. Accuracy and Elapsed Time

The relationship between elapsed time and accuracy also take into consideration. Sometime, shorter time does not mean more accurate. The time affects the accuracy by processing as much as possible data.

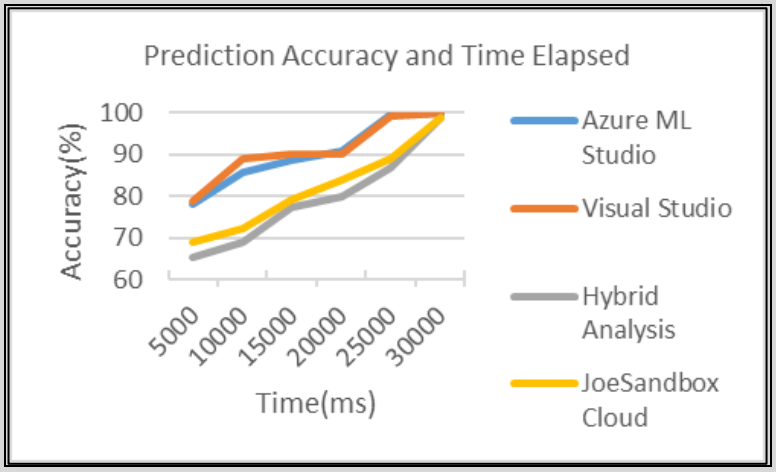


Figure 4.24: Relationship between accuracy and elapsed time

4.4 SUMMARY

This chapter discuss about the implementation and testing of the spam detection by using machine learning research that has been carried out. So, all of the module in this application are tested to generate final output to make sure that the model is correctly created. This is the crucial part of the project development.

CHAPTER 5

CONCLUSION

5.1 EXPECTED RESULT

As state before, supervised ML is able to separate messages and classified the correct categories efficiently. It also able to score the model and weight them successfully. For instances, Gmail's interface is using the algorithm based on ML program to keep their users' inbox free of spam messages.

5.2 LIMITATION AND CONSTRAINT

Based on the result of this project, only text (messages) can be classified and score instead of domain name and email address. This project only focus on filtering, analysing and classifying message and do not blocking them.

5.3 SUGGESTION AND IMPROVEMENT

Some suggestion that can be applied to this project is to widen its use to not just classifying only text message format. So, an improvement can be made to leverage the use of this project so that it can filter, analyse, classify and score model not limited to just text message but including any other format such as domain name. In other to get the most accurate result of classification, these improvements should be made.

5.4 SUMMARY

From this project, it can be concluded that machine learning algorithm is one of the important part in order to create spam detection application. To make it more efficient, improvement need to be implemented in future.

REFERENCES

- https://www.geeksforgeeks.org/
- Kaggle: Your Machine Learning and Data Science Community
- machine learning - Google Scholar