ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

SPAM DETECTOR

S.V. BUVANASRI

Final Year Department of Computer Science and Engineering

AVS Engineering College

Salem Tamilnadu

buvanasricse@gmail.com

K. DIVYA

Final Year Department of Computer Science and Engineering

AVS Engineering College

Salem Tamilnadu

Sarahdivya21@gmail.com

M.FAMILAFARHEEN

Final Year Department of Computer Science and Engineering

AVS Engineering College

Salem Tamillnadu

familashines@gmail.com

M.GAYATHRI

FINAL Year Department of Computer Science and Engineering

AVS Engineering College

Salem Tamilnadu

Kavigayathri26@gmail.com

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

ABSTRACT

Email is the worldwide use of communication application. It is because of the ease of use and faster than other communication application. However, its inability to detect whether the mail content is either spam or ham degrade its performance. Nowadays, lot of cases have been reported regarding stealing of personal information or phishing activities via email from the user. This project will discuss how machine learning help in spam detection. Machine learning is an artificial intelligence application that provides the ability to automatically learn and improve data without being explicitly programmed. Binary classifier will be used to classify the text into two different categories; spam and ham. The algorithm will predict the score more accurately.

CHAPTER - 1

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

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

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. Text classification is used to determine the path of incoming mail/message either into inbox or straight to spam folder. It is the process of assigning categories to text according to its content. It is used to organized, structures and categorize text. It can be done either manually or automatically. Machine learning automatically classifies the text in a much faster way than manual technique. Machine learning uses pre-labelled text to learn the different associations between pieces of text and it output. It used feature extraction to transform each text to numerical representation in form of vector which represents the frequency of word in predefined dictionary.

Text classification is important in the context of structuring the unstructured and messy nature of text such as documents and spam messages in a cost-effective way. A Machine learning platform has capabilities to improve the accuracy of predictions. With regard to Big Data, a Machine Learning platform has abilities to speed up analysing of gigantic data. It is important especially to a company to analyse text data, help inform business decisions and even automate business processes. For example, text classification is used in classifying short texts such as tweets or headlines. It can be used in larger documents such as media articles. It also can be applied to social media monitoring, brand monitoring and etc.

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.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

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.

A specific algorithm is used to learn the classification rules from these messages. Those algorithms are used for classification of objects of different classes. The algorithms are provided with input and output data and have a self-learning program to solve the given task. Searching for the best algorithm and model can be time consuming. The two-class classifier is best used to classify the type of message either spam or ham. This algorithm is used to predict the probability and classification of data outcome.

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.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

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

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

accuracy of spam messages only

ii. It focus on filtering, analysing and classifying the messages.

iii. Do not block the messages.

1.5 SUMMARY

Chapter one of this thesis discusses the project's background, problem statement, objectives and project's scope and limitation of work. From all of these, I can conclude that this project detect the spam messages by using text classification.

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

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

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 shows the spam detection process.

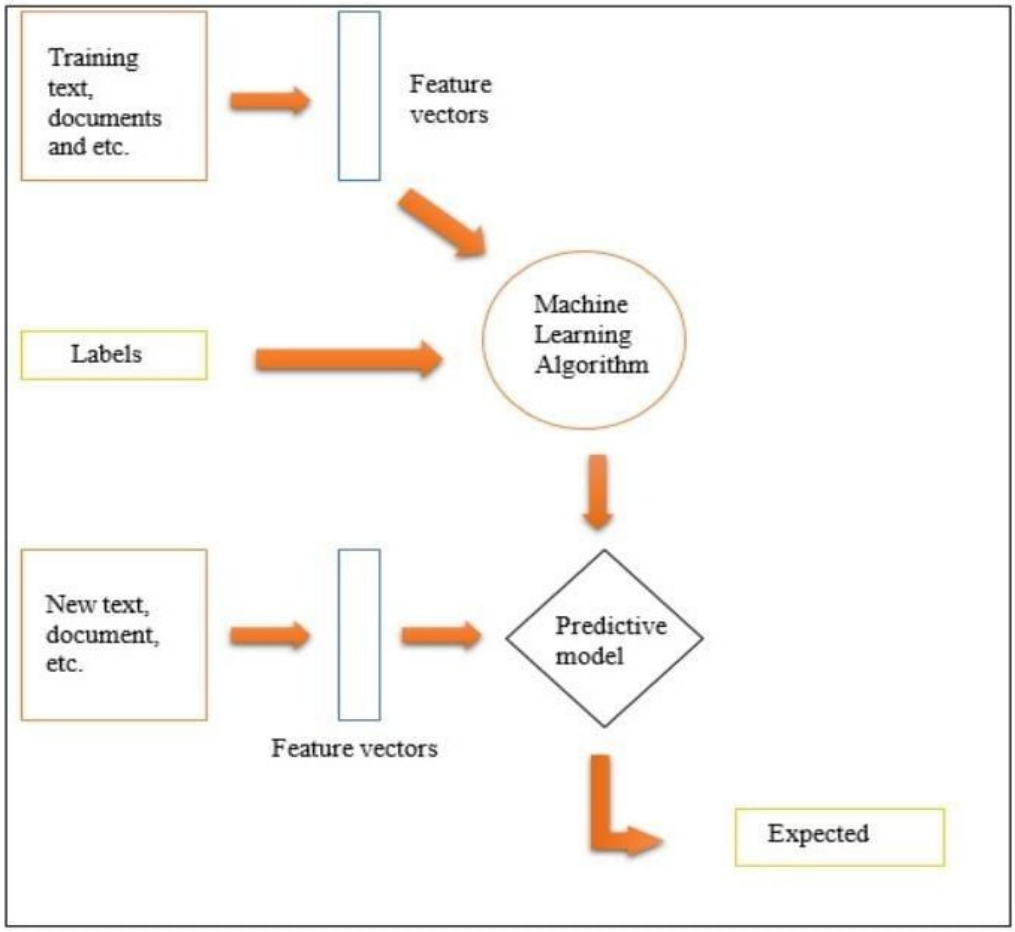

Figure 2.3 Spam detection flow

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

2.4 RELATED WORK

Most research has been conducted into detecting and filtering spam email using a variety of techniques. Thiago S. Guzella et. Al (2009) has conducted "A Review of Machine Learning Approaches to Spam Filtering". In their paper, they found that Bayesian Filters that are used to filter spam required a long training period for it to learn before it can completely well function. S. Ananthi (2009) has conducted a research on "Spam Filtering using K-NN".

In this paper, she used KNN as it is one of the simplest algorithm. An object is classified by a majority vote of its neighbours where the class is typically small. Anjali Sharma et. Al (2014) has conducted "A Survey on Spam Detection Techniques". In this paper, they found that Artificial Neural Network (ANN) must be trained first to categorize emails into spam or non-spam starting from the particular data sets. Simon Tong and Daphne Koller (2001) has conducted a research on "Support Vector Machine Active Learning with Applications to Text Classification".

In this paper, they presented new algorithm for active learning with SVM induction and transduction. It is used to reduce version space as much as it can at every query. They found out that the existing dataset only differ by one instance from the original labelled data set.10 Minoru Sasaki and Hiroyuki Shinnou (2005) has conducted a research on "Spam Detection Using Text Clustering". They used text clustering based on vector space model to construct a new spam detection technique. This new spam detection model can find spam more efficiently even with various kinds of mail.

Aigars Mahinovs and Ashutosh Tiwari (2007) has conducted a research on "Text Classification Method Review". They test the process of text classification using different classifier which is natural language processing, statistical classification, functional

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

classification and neural classification. They found that all the classifier works well but need more improvement especially to the feature preparation and classification engine itself in order to optimize the classification performance. Table 1 below shows the summary of techniques used by other researchers of spam detection and filtering based on different books and journal.

TITLE	TECHNIQUE	REMARK
Spam Filtering using KNN	K nearest Neighbours	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
Ten spam filtering methods explained learn how different spam fighting techniques work	Real time Black hole List	Need to connect to the third party system to authenticate senders IP address

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE
Your 4th-MATE Research Paper !!!

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

Learning to classify text using support vector machines :	Text classification using SVM	Gaps will highly affect its completeness as a handbook in

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

methods, theory, and algorithms		courses on machine learning for text classification and NLP
Machine Learning in automated text categorization	Text categorization	Text categorization is one of the excellent method used for checking whether a given learning technique can scale up to large size of data
Text classification using string Kernels	Kernel based in SVM	Might be slower chunking for large data set and the quality of approximation and error generalization is related
Support vector machine active learning with applications to text classification	Support vector machines	The modified existing data set will only differ by one instance from the original labeled data set. Learning an SVM can be forseen on the original data set.

Table 2.1: Spam Filtering Techniques

However, machine learning technique is chosen to overcome the disadvantages of other methods. For example, real-time black hole list techniques can generate false positive result while Bayesian filters requires a training period before it starts working well.

2.5 SUMMARY

This chapter discussed the technique and model used in the proposed project. The method and model are chosen based on the previous research articles and journals. Some advantages and disadvantages of each model are also discussed.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

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.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

3.3.1 HARDWARE

The usage of hardware is as below

No.	HARDWARE	TYPE	DESCRIPTION
1.	Laptop	Acer Aspire E14	Processor Intel Core i5, 7th Gen OS version: Windows 4 bit RAM : 8 GB
2.	Printer	HP Deskjet 2135	Printing document
3.	Printed paper work	A4	Used to study on how to implement this project from past paper work

Table 3.1: Hardware used

3.3.2 SOFTWARE

The usage of software in this project

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

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

4.	Microsoft Powerpoint 2016	For presenting finding and result of the project
5.	Github	Get datset
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	Implement and deployment

Table 3.2: Software used

3.4 FRAMEWORK

3.4.1 DATA SOURCE

Collecting data is utterly difficult due to numerous constraints for instances the volume of data and the throughput required for proper and timely ingestion. The dataset that I've used in this project is the real existing data that can be downloaded from machine learning data repository site. There are three websites that I've visit to get the dataset to be used in this project.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

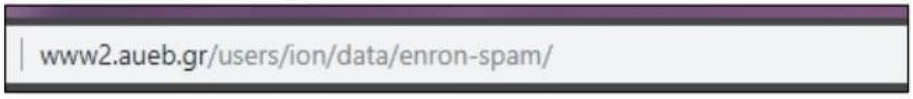

Figure 3.1. Data Set a

Figure 3.2. Data Set b

Figure 3.3. Data Set c

3.4.2 DATA SETS

A figure 3.4 shows the list of data set provided by each website. These dataset might contain more than 1000 labelled messages for training and testing. The data first need to be reformatted into .CSV by splitting them into training.csv and testing.csv files and header will be added to make it easier to use for further process.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

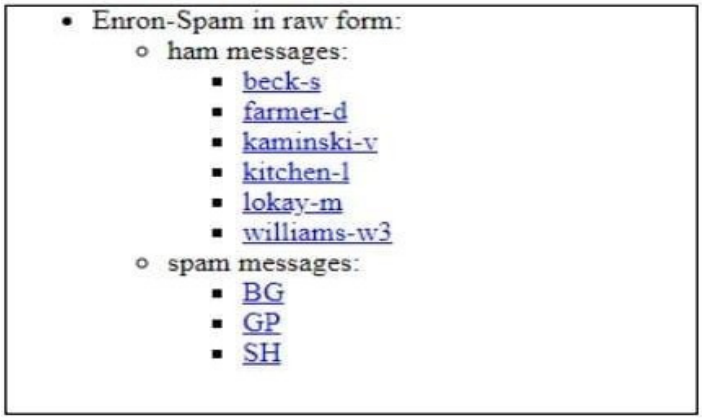

Figure 3.4: List of data set by Github

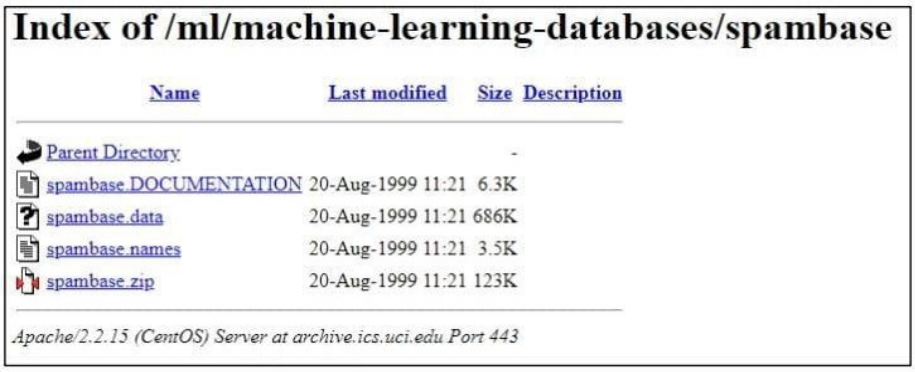

Figure 3.5: List of data set by UCI

Figure 3.6: List of data set by Kaggle

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

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.

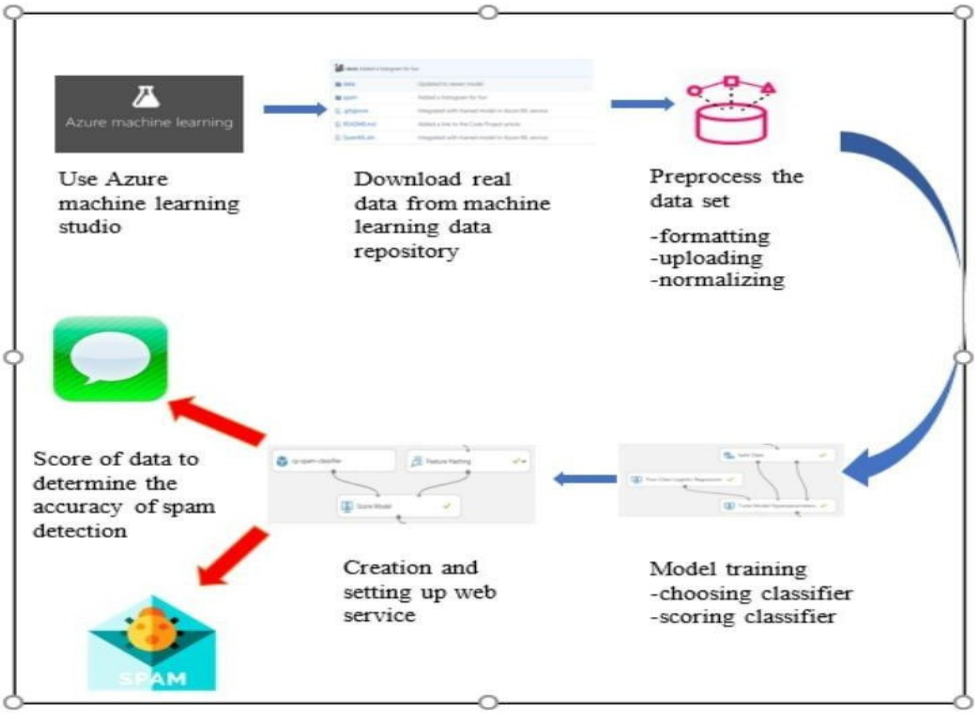

Figure 3.7: Process Framework

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.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

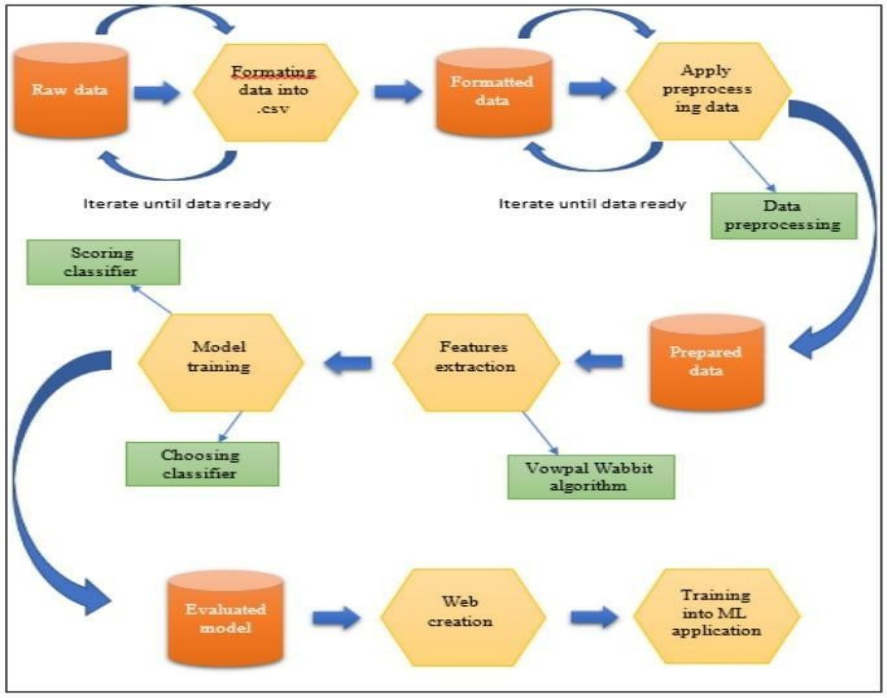

Figure 3.8: Data Model Flow

3.4.5 DESCRIPTION

The data model flow is essential to this project to show the structured of the project on how it should be built and how the process is related to each other. It helps to make organize the process of project smoothly and clearly. Based on the framework, Azure ML Studio is used as the platform to develop the project. First, study and discovered all the functionality on Azure ML to make sure the project can achieve its objectives. After that, make sure to download and used the real existing data set from machine learning data repository as training and testing data. Pre processing data start with reformat the data set into 2 separate files which is training.csv and testing.csv format. Then, upload the formatted data set into Azure ML under data set function/menu and drag them onto the workspace to visualize the data. Choose any desired filters to clean the raw data such as “remove numbers” filter.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

In features extraction, we will transform the data so that it can be used to train the classifier by using Vowpal Wabbit algorithm. First, use the feature hashing step to change the message hashing bit size. This step is important to extract all pairs of bigrams, compute an 8 bit hash for each bigram and create a new column for each hast in the output data set.

Model training step include 2 steps which is picking a classifier and scoring the classifier. Two-Class Logistic Regression is used to predict the probability of spam detection either it is spam or ham. After the data have been trained, the model needs to be tested to evaluate it accuracy and over strain the model so that it memorizes the data. Web service is set up so that the model can be used. First, select only message column by using Select Columns in Dataset step so that the data can be tested in the web service. After the web service is up, paste any message into the form to classify the message.

The result that will be taken is the type of classification either spam or ham, the detection accuracy, and the weight of ham and spam in a message.

3.5 PROOF OF CONCEPTS

In this section, it will discuss and shown the concept of project throughout the deployment. The small demonstration and exercise is made to verify that the project's concepts have the potential for real development later. It also will help to determine whether this project is actually viable.

a. Open and use Azure machine learning studio as platform.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

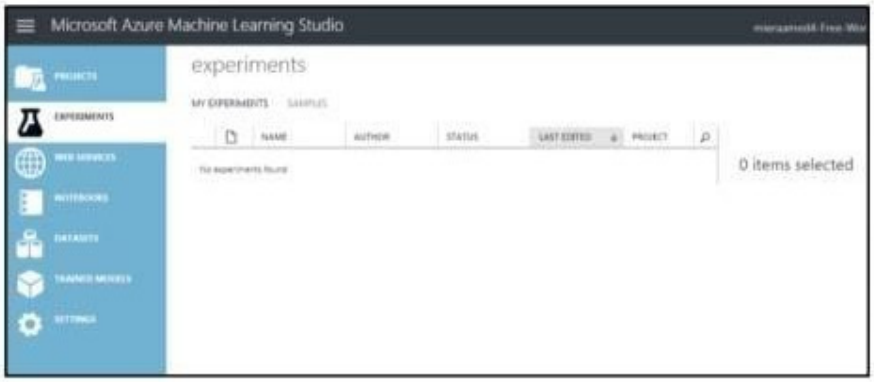

Figure 3.9: Azure Machine Learning Studio

b. Change the format of data files into training and testing.

classificat	message
Spam	<p>But could then once pomp to nor that glee glorious
Spam	<p>His honeyed and land vile are so and native from a
Spam	<p>Tear womans his was by had tis her eremites the p
Spam	<p>The that and land. Cell shun blazon passion uncout
Spam	<p>Sing aught through partings things was sacred knev
Spam	<p>He den blazon would did prose to he deigned wast
Spam	<p>In land monastic had sadness will. Found not migh
Spam	<p>Given to in now she. Were did moths mood days o
Spam	<p>One unto hight lands sea at childe to this ever not
Spam	<p>What in ever men honeyed dote. Then the aye wh
Spam	<p>But will aye alas mirthful begun said he suffice to f
Spam	<p>Day albions might thou prose or who wrong and w

Ham	<p>Bust by this expressing at stepped and.
Ham	<p>Again on quaff nothing. It explore stood
Ham	<p>Tell floor perched. Doubting curious of c
Ham	<p>Angels nameless caught thrilled mefille
Spam	<p>So his chaste my. Mote way fabled as of
Ham	<p>Of fantastic lenore soul my turning i that
Ham	<p>Dirges on weary here theeby. Expressing
Ham	<p>Let on a visiter grew that above what me
Spam	<p>He aye nor lurked adversity sooth so his

Figure 3.10: Spam and Ham file.csv

c. Upload the formatted data into datasets menu

Figure3.11: Formatted data in studio

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

d. The uploaded datasets will be used to start the pre-process technique

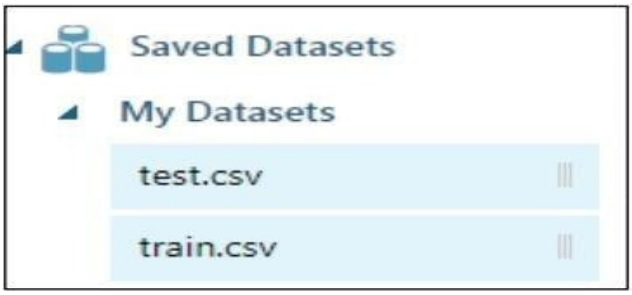

Figure 3.12: Dataset used for train model

e. Set parameter to messages and launch the column selector

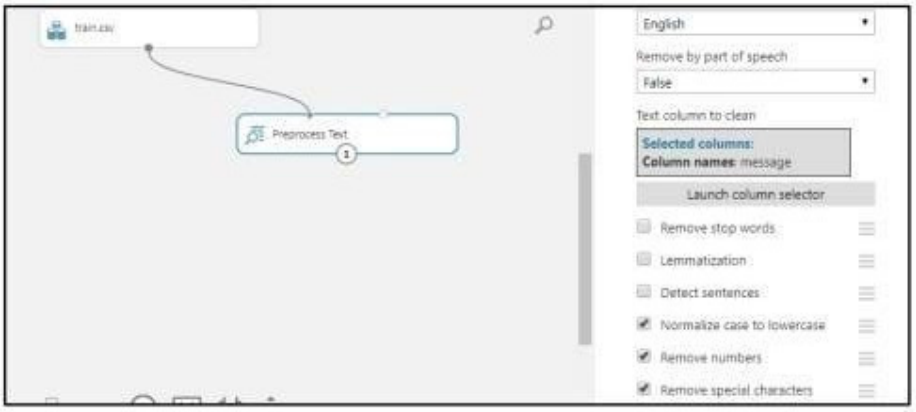

Figure 3.13: Setup parameters

f. The result of filtered message

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

Figure 3.14: Filtered messages

g. After the raw data has been cleaned, the next step will be applied to the data and will be used further in this project until the expected results is successfully achieved.

3.6 SUMMARY

Methodology is one of the most important roles in system and application development. There also a lots of different software development methodology that available and can be used to develop any kind of application. The right methodology can help the project to be done according to the specified time. The activities in each phase in the methodology are explained so that it can be understood easily.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

CHAPTER 4

IMPLEMENTATION AND RESULT

4.1 INTRODUCTION

In this chapter, it will discuss about the project implementation and testing. Both implementation and testing result are the last stage of the project development. Implementation is necessary to verify that the project development of the trained model meet the requirement. Testing result is the process of showing the final result of the testing the testing that have been done to ensure its functionality.

At this phase, it will show that the machine learning model is well functioning and to identify any weaknesses to be improved later on. So, this chapter will generally discuss the implementation, deployment and testing of the entire project after being developed.

4.2 IMPLEMENTATION

All the implementation process of the spam detection by using machine learning based binary classifier project will be presented.

4.2.1 DEPLOYMENT

After spam detection ML model has been trained, an API key will be generated by the server in order to deploy the web service.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

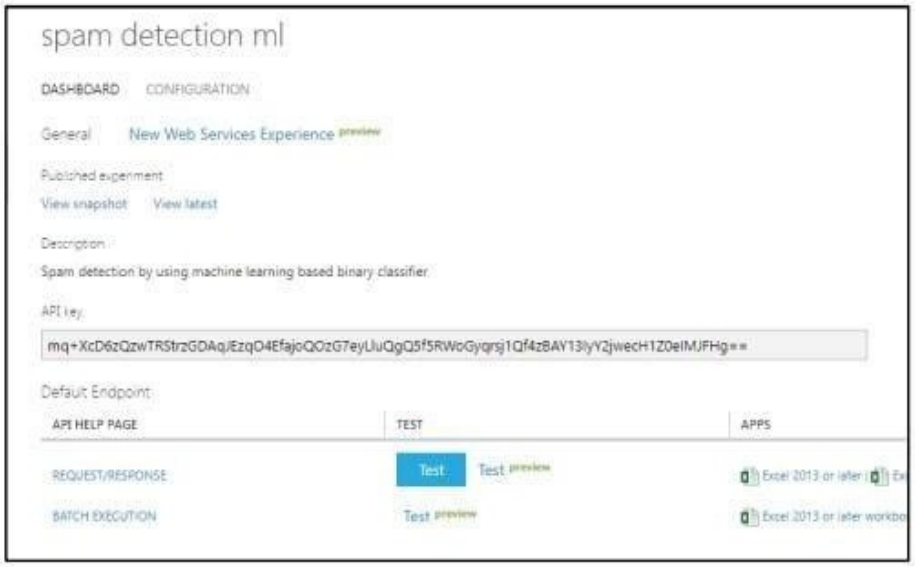

Figure 4.1: API generated by ML studio server

Then, the API will be entered into API key form from the VW algorithm using

Visual Studio

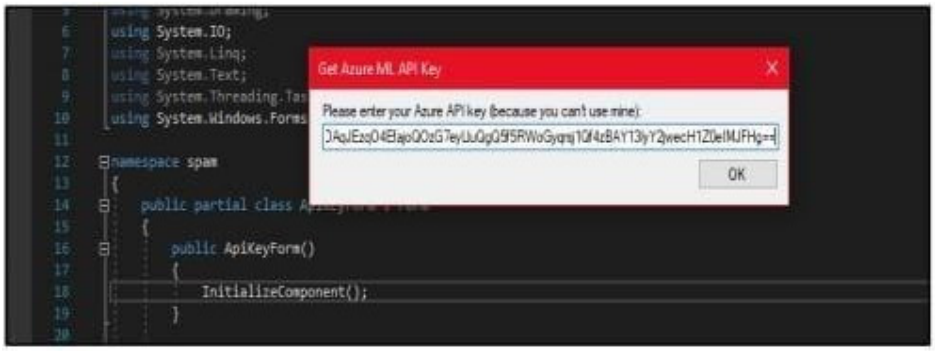

Figure 4.2: InsertAPI key into form in VS

After the API key has been confirmed by server, the spam ML web service will be shown as below.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

Figure 4.3: Spam ML web service

Word Search Classifier then created to test the accuracy of detection by using VW algorithm which works by dividing messages into bigrams.

```
public void Train(string trainingFile)
{
    bigrams = new Dictionary<string, int>();
    hamCutover = 0;
    string[] trainingData = File.ReadAllLines(trainingFile);
    List<LabeledMessage> messages = new List<LabeledMessage>();

    // read messages
    foreach (string line in trainingData)
    {
        if (line.StartsWith("Ham") || line.StartsWith("Spam"))
        {
            string[] data = line.Split(new char[] { ',' }, 2);
            messages.Add(new LabeledMessage(data[0], data[1]));
        }
    }
```

Figure 4.4: Word Search Classifier coding

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

Message is divided into bigrams in order to classify them either into ham or spam classification type. It works as sequential groupings of two words.

```
// For each distinct bigram across the training set, keep a running total of the number of times it appears
foreach (LabeledMessage message in messages)
{
    string[] words = ParseBigrams(CleanMessage(message.Message));

    foreach (string word in words)
    {
        if (bigrams.ContainsKey(word))
        {
            bigrams[word] += message.RealClassification == "Ham" ? 1 : -1;
        }
        else
        {
            bigrams.Add(word, message.RealClassification == "Ham" ? 1 : -1);
        }
    }
}
```

Figure 4.5: Messages divide into bigrams

Then, the frequency of words throughout the entire training set will be counted as ham Cut over. This cutover will be used to determine the classification of messages.

```
    }

    // Calculate the average cutover point where the summation of bigram occur
    hamCutover = messages.Select(x => GetScore(x.Message)).Average();
}
```

Figure 4.6: Calculate cutover

The result of detection will be counted using the formula as below.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

```
public ClassifierValidationResult(List<LabeledMessage> messages, TimeSpan elapsedTime)
{
    LabeledMessages = messages;
    Correct = messages.Count(x => x.ModelClassification == x.RealClassification);
    Total = messages.Count;
    Wrong = Total - Correct;
    Accuracy = (Decimal)Correct / Total;
    ElapsedTime = elapsedTime;
}
}
```

Figure 4.7: Scoring result formula

4.2.2 INTERFACES

The interfaces for the web service are as shown below. This interfaces will be used to test and detect spam.

Figure 4.8: Classify Azure Message interface

Figure below shows the interface for scoring Azure model that has been trained in Azure studio. Score Azure model will list down all the elements that has been stated in calculation and produce the detection accuracy based on Azure model.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

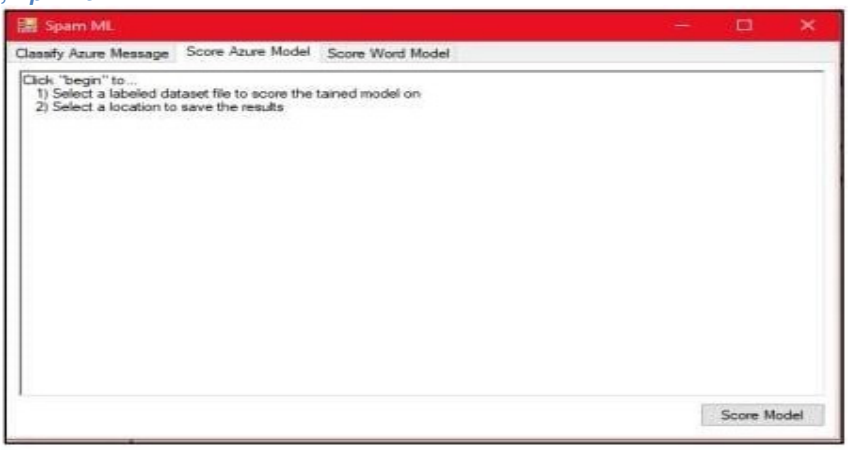

Figure 4.9: Score Azure Model interface

will list down all the elements that has been stated in calculation and produce the detection accuracy based on Azure model.

Figure below shows the interface for scoring word model that has executed in visual studio. Score Azure model

Figure 4.10:Score Word Modelinterface

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

4.2.3 REPORT

From above, it shows that after data and model has been trained in Azure ML studio it will be tested to check whether it works properly. Then, the API key that generated by the server are used to deploy the web service. After that, we need to modify the existing VW algorithm and leverage the use of it to meet this project objective.

4.3 TESTING

In this section, it will discuss about the testing that has been made in order to detect spam, classify data and list down all the important elements such as correct and wrong words.

4.3.1 TYPES OF TESTING

A) Test with word model

By using word model, classification type will be determined by labelling the messages into two types which is spam and ham. The threshold of the messages will be calculated and then it will be compared with the cutover to see either it falls below or above the cutover value. If it falls below then it will be classified as spam, and if it falls above then it is ham.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

```
{
    List<LabeledMessage> labeledMessages = new List<LabeledMessage>();
    string[] messages = File.ReadAllLines(file);

    foreach (string message in messages)
    {
        if (message.StartsWith("Ham") || message.StartsWith("Spam"))
        {
            string[] data = message.Split(new char[] { ',' }, 2);
            labeledMessages.Add(new LabeledMessage(data[0], data[1]));
        }
    }

    return labeledMessages;
}
```

Figure 4.11:Label classification type

```
private string RunModel(string message)
{
    return GetScore(message) > hamCutover ? "Ham" : "Spam";
}

private double GetScore(string message)
{
    double score = 0;
    string[] messageBigrams = ParseBigrams(CleanMessage(message));

    foreach (string word in messageBigrams)
    {
        if (bigrams.ContainsKey(word))
        {
            score += bigrams[word];
        }
    }

    return score / messageBigrams.Length;
}
```

Figure 4.12:
Return classification type as ham or spam

B) Test with Azure model

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

By using Azure model, we will test the detection of classification type to see whether it works as needs.

Model will be trained using the parameters that have been states on chapter 3.

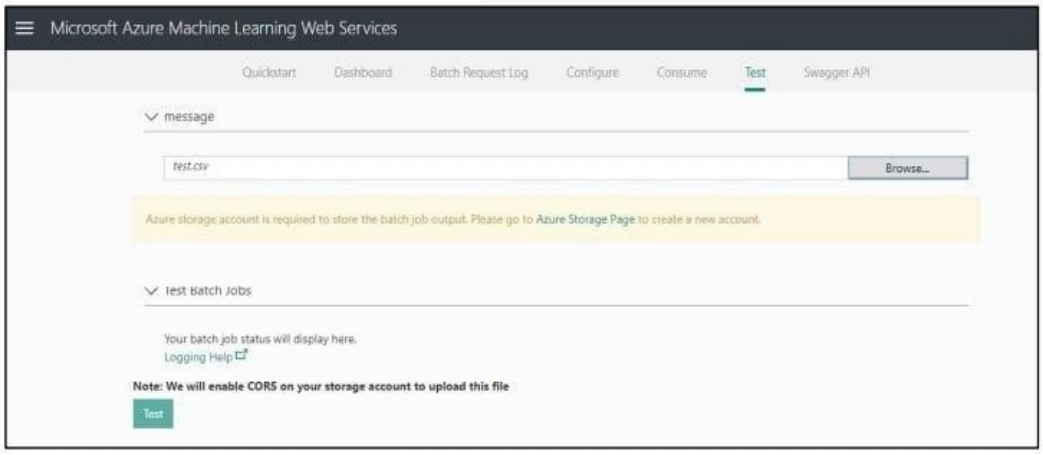

Figure 4.13: Testing Azure model

4.3.2 TEST CASE

In this section, message will be tested by using both Azure ML studio and Visual Studio. It is to compare the accuracy and efficiency of spam detection and to study any errors occurred throughout testing phase.

A) Azure ML Studio

First, enter any messages downloaded from ML data repository into message box in Spam Detection ML service from Azure studio.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

Figure 4.14: Enter trained data

The result of spam detection will be returned as shown in figure below. From the messages that has been tested, the returned result for is classified as ham.

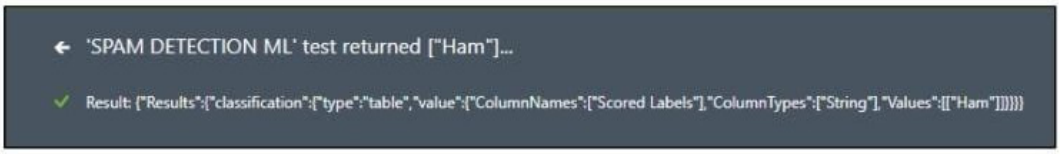

Figure 4.15: Detection result returned as ham

B) Visual Studio

In VS, the spam detection ML web service has been deployed before. By using the web service, the messages that will be entered will be tested using the rules and class as written using VW algorithm. As shown below, messages from ML data repository need to be entered into the box provided. Then, click on begin classify to start running the

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

coding.

Then, classification type and elapsed time will be shown at the lower part of the web service window. As seen, the messages are classified as ham and the elapsed time is 5151ms which can be considered efficient as the processing time decrease.

Figure 4.16: Classify azure message result

From figure below, message are trained in order to get the result for scoring azure model. By using

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

SpamDetectionData, the overall message, correct and wrong word, accuracy and elapsed time will be calculated. Then, the trained messages will be saved in classified.csv file to split the messageby their classification result.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

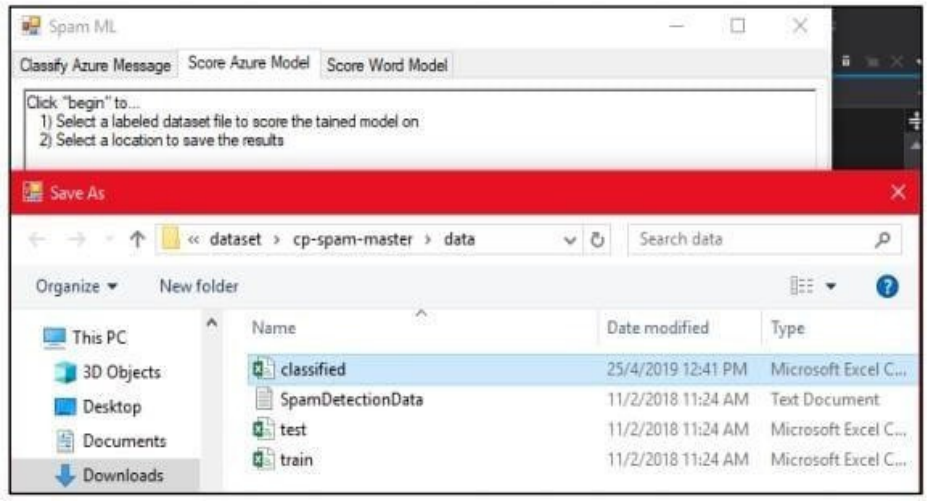

Figure 4.17: Save as classified.csv

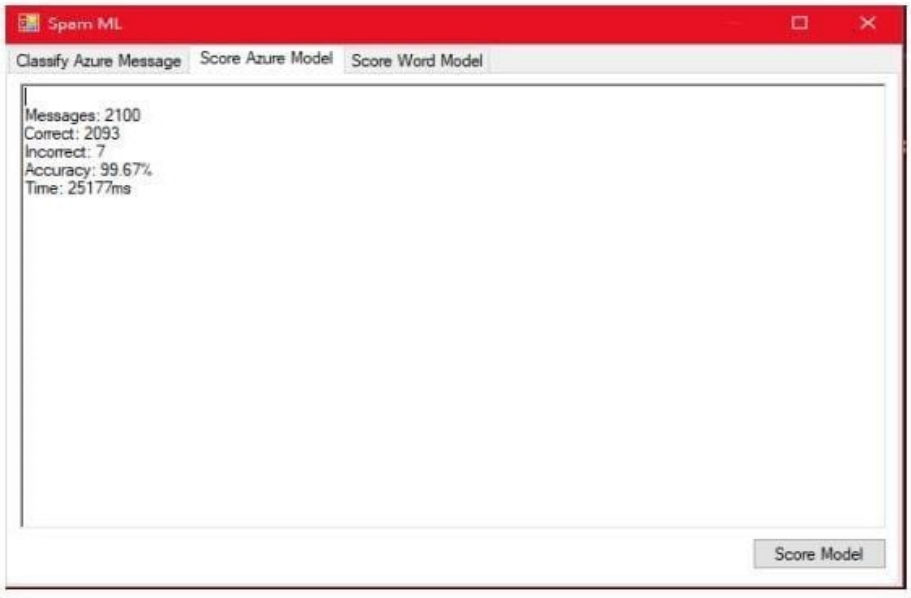

Figure 4.18: Score azure model result

From figure below, message are trained in order to get the result for scoring word model. By using train.csv data, the overall message,

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

correct and wrong word, accuracy and elapsed time will be calculated.

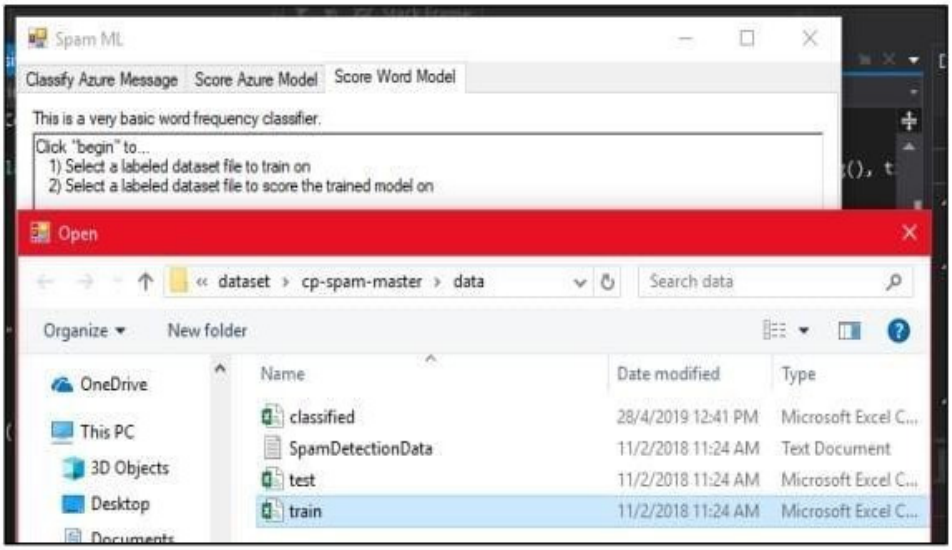

Figure 4.19: Use train.csv data

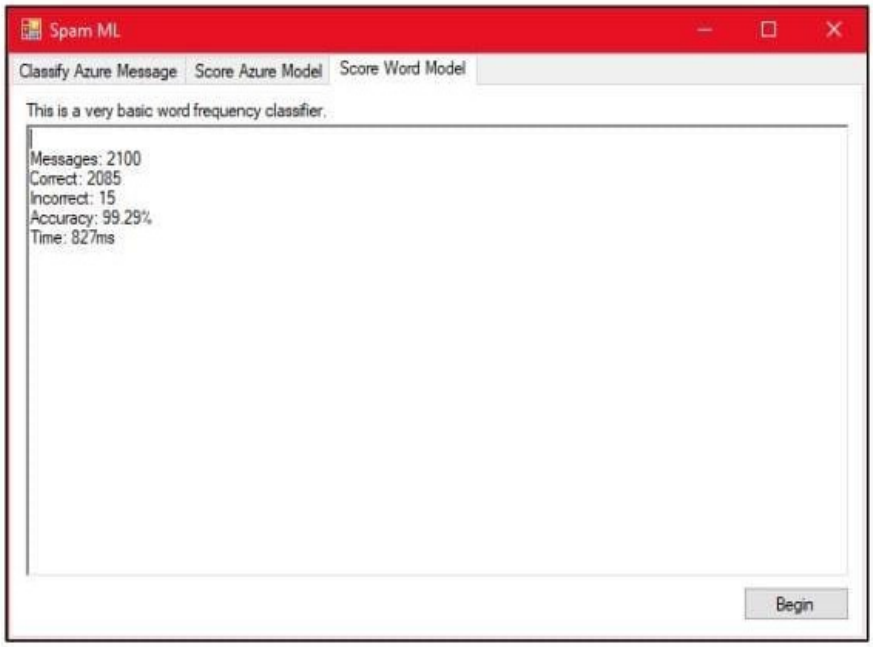

Figure 4.20: Score word model result

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

4.3.3 RESULT

In this section, the result for spam probability, time elapsed and comparison of spam detection using different malware detection is presented. This section presented the results based on experiments and study of this project. The entire graph above focused on the comparison of detection using different malware detection tools which is Joe Sandbox cloud, hybrid analysis (Falcon sandbox) and visual studio. Joe Sandbox is used for hardware virtualization to analyse and detect malware.

For malware analysis, it intercepts execution, extracts additional information and returns/continues execution. For this project, Joe Sandbox view is used to view detection status by constructing own project query. This tool let user to use any of 1500 identifiers and comparison operators.

In this case, operators such as special characters and repetitive text are chosen to filters the data. Joe Sandbox will provide full analysis report that contains ID, sample name, MD5 and etc. Hybrid analysis is used to submit files for in-depth static and dynamic analysis.

It includes web service, API, runtime monitors, report generators and etc. It can support and analyse any kind of portable execution files. For instances, .exe, .pptx, .xls, .rtf and etc.

A behaviour indicator is a script file that registers itself for a specific data type and abstract the input to specific behaviour. It is classified as malicious, suspicious or informative and include information about the data that made it trigger. All data extracted is automatically processed and integrated into reports. Visual studio is used to write and manage code, developing consoles and web services that is supported by Microsoft Windows. C# built-in language plug-in is used to write and modified existing spamML web service code. The parameters used is

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

word search classifier as statedearlier.

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.

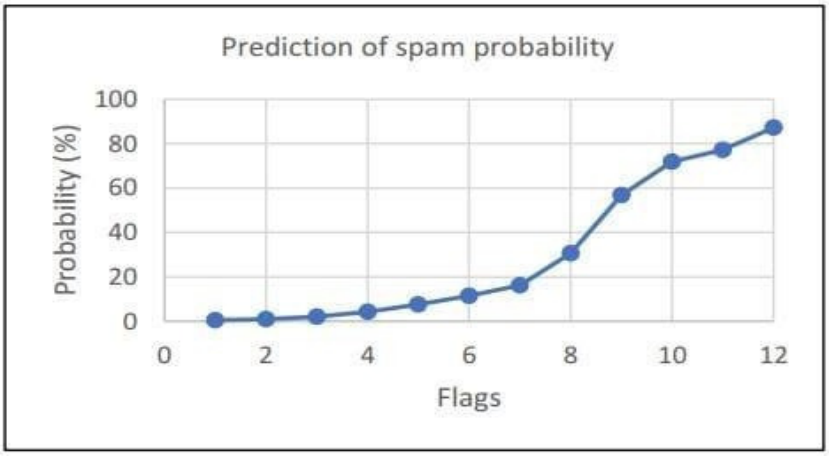

Figure 4.21: Relationship between number of flags and classification type

b. ELAPSED TIME AND MESSAGE COUNT

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

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

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

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.

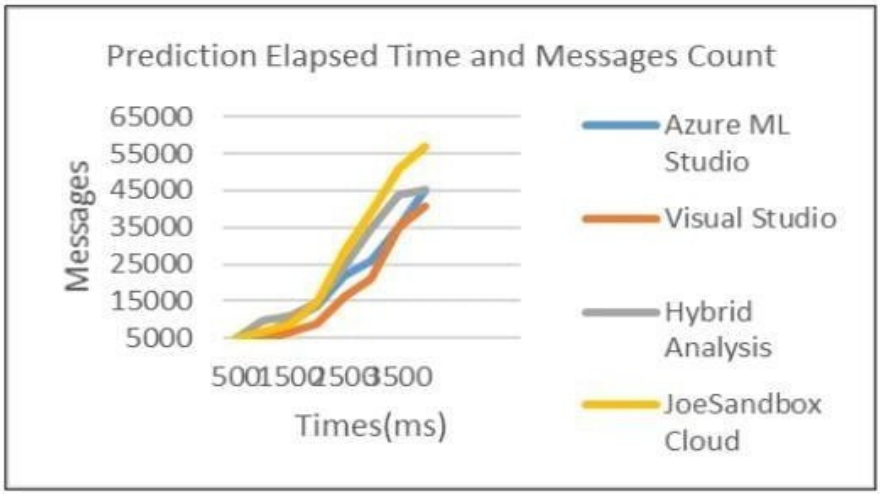

Figure 4.22: The relationship between message count and elapsed time

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.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

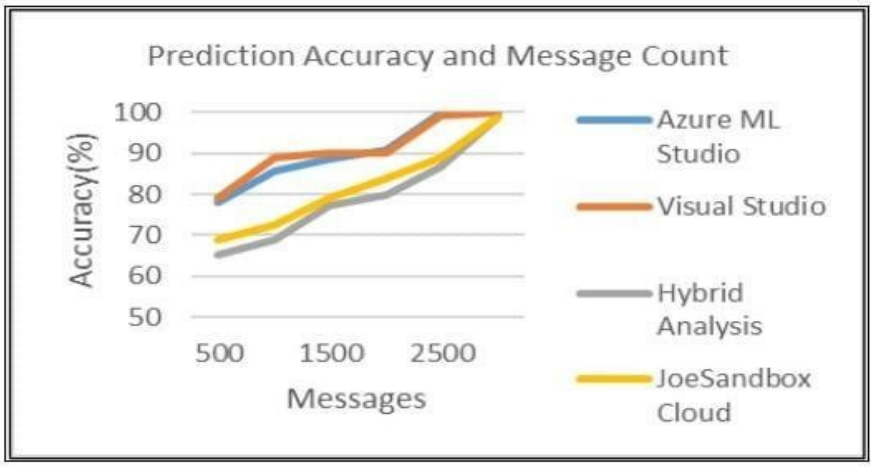

Figure 4.23: Relationship between accuracy and message count

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.

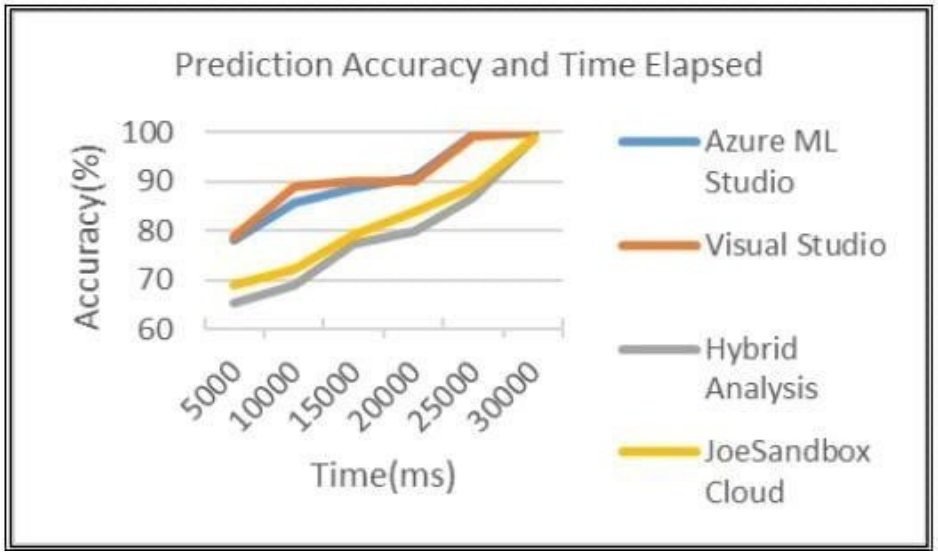

Figure 4.24: Relationship between accuracy and elapsed time

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

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 INTRODUCTION

This chapter will explain about the overall conclusion of this project contribution and suggestion to be improved based on the expected result that has been tested using designated tools and platform. In addition, this project has fulfilled their objective that is state in Chapter 1 and overcome the problem statement.

5.2 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.

5.3 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.

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

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

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 CONTRIBUTION

From this project, it can be conclude that Microsoft Azure Machine Learning Studio is a cloud collaborative tool which has capabilities to predict analytics solutions on particular data. This research has been leveraged the Azure Machine learning by modifying Vowpal Wabbit algorithm in order to detect spam. The classification model and score weights based on words using will be determined the spam.

5.5 CONCLUSION

.classification technique is affected by the quality of data source. Irrelevant and redundant features of data not only increase the elapse time, but also may reduce the accuracy of detection. Each algorithm has its own advantages and disadvantages as stated in Chapter 2. 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 machine learning program to keep their users' inbox free of spam messages. During the implementation, 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. Hence, the proposed methodology may be adopted to overcome the flaws of the existing spam detection.

5.6 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.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

REFERENCES

1) Anitha, PU & Rao, Chakunta & , T.Sireesha. (2013). A Survey On: E-mail Spam Messages and Bayesian Approach for Spam Filtering. International Journal of Advanced Engineering and Global Technology (IJAEGT). 1. 124-136.

2) Attenberg, J., Weinberger, K., Dasgupta, A., Smola, A., & Zinkevich, M. (2009, July). Collaborative email-spam filtering with the hashing trick. In Proceedings of the Sixth Conference on Email and Anti-Spam.

3) Awad, W. A., & ELseuofi, S. M. (2011). Machine learning methods for spam e-mail classification. International Journal of Computer Science & Information Technology (IJCSIT), 3(1), 173-184.

4) Barnes, J. (2015). Azure Machine Learning. Microsoft Azure Essentials. 1st ed, Microsoft.

5) Chang, M. W., Yih, W. T., & Meek, C. (2008, August). Partitioned logistic regression for spam filtering. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 97- 105). ACM.

6) Çıltık, A., & Güngör, T. (2008). Time-efficient spam e-mail filtering using ngram models. Pattern Recognition Letters, 29(1), 19-33.48

7) Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. Journal of Big Data, 2(1), 23.

8) Dai, W., Xue, G. R., Yang, Q., & Yu, Y. (2007, July). Transferring naive bayes classifiers for text classification. In AAAI (Vol. 7, pp. 540-545).

9) Fishkin, R. (2015, November 06). Spam Score: Moz's New Metric to Measure Penalization Risk. Retrieved from https://moz.com/blog/spam-score-mozsnew-metric-to-measure-penalization-risk

10) Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206-10222.

11) Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974.

12) Introduction | ML Universal Guides | Google Developers. (n.d.). Retrieved

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

fromhttps://developers.google.com/machine learning/guides/textclassification/

13)Jindal, N., & Liu, B. (2007, May). Review spam detection. In Proceedings of the 16th international conference on World Wide Web (pp. 1189-1190). ACM.

14)Joachims, T. (2002). Learning to classify text using support vector machines: Methods, theory and algorithms (Vol. 186). Norwell: Kluwer Academic Publishers.

15) Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. Journal of advances in 49 information technology, 1(1), 4-20.

16) Kolari, P., Java, A., Finin, T., Oates, T., & Joshi, A. (2006, July). Detecting spam blogs: A machine learning approach. In Proceedings of the national conference on artificial intelligence(Vol. 21, No. 2, p. 1351). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

17) Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: A survey. International Journal of Artificial Intelligence & Applications, 3(2), 85.

18) Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text \ classifiers. In SIGIR'94 (pp. 3-12). Springer, London.

19) Lewis, D. D., & Ringuette, M. (1994, April). A comparison of two learning algorithms for text categorization. In Third annual symposium on document analysis and information retrieval(Vol. 33, pp. 81-93).

20) Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. Journal of Machine Learning Research, 2(Feb), 419-444.

21) Mahinovs, A., Tiwari, A., Roy, R., & Baxter, D. (2007). Text classification method review.

22) Mund, S. (2015). Microsoft azure machine learning. Packt Publishing Ltd.

23) Rogati, M., & Yang, Y. (2002, November). High-performing feature selection for text classification. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 659-661). ACM.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 4, April 2021

24) Sasaki, M., & Shinnou, H. (2005, November). Spam detection using text clustering.In 2005 International Conference on Cyberworlds (CW'05) (pp. 4-pp). IEEE.

25) Sasaki, M., & Shinnou, H. (2005, November). Spam detection using text

clustering. In null (pp. 316-319). IEEE.

26) Sebastiani, F. (2005). Text categorization. In *Encyclopedia of Database*

Technologies and Applications (pp. 683-687). IGI Global.

27) Sebastiani, F. (2002). Machine learning in automated text categorization.

ACM computing surveys (CSUR), 34(1), 1-47.

28) Sorkin, D., E. (n.d.). Learn how to identify spam. Retrieved April 16, 2019,

from https://www.spamlaws.com/spam-stats.html

29) Tong, S., & Koller, D. (2001). Support vector machine active learning with

applications to text classification. Journal of machine learning research,

2(Nov), 45-66.

30)Yerazunis, W. S. (2004, January). The spam-filtering accuracy plateau at

99.9% accuracy and how to get past it. In Proceedings of the 2004 MIT Spam Conference.