Annexure – 1

**SPAM EMAIL DETECTION**

***A Project Report submitted***

***in partial fulfillment of the requirements***

***for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

***In***

**COMPUTER SCIENCE & ENGINEERING**

***By***

1. VISSAKODETI MALLIKA 3. S.V. SRAVANA SUREKHA

2. PENMATSA SRI RAMYA JYOTHI 4. VARDHANAPU KEERTHI

***Under the esteemed guidance of***

**Mr. T Kesava (qualification)**

**Designation**



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**SHRI VISHNU ENGINEERING COLLEGE FOR WOMEN(A)**

**(Approved by AICTE, Accredited by NBA & NAAC, Affiliated to JNTU Kakinada)**

**BHIMAVARAM – 534 202**

**2020 – 2021**

Annexure – 2

**SHRI VISHNU ENGINEERING COLLEGE FOR WOMEN(A)**

**(Approved by AICTE, Accredited by NBA & NAAC, Affiliated to JNTU Kakinada)**

**BHIMAVARAM – 534 202**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**



**CERTIFICATE**

*This is to certify that the project entitled “****Spam Email Detection****”, is being submitted by* ***Keerthi Vardhanapu*** *bearing the* ***Regd. No. 17B01A05G9*** *in partial fulfillment of the requirements for the award of the degree of “****Bachelor of Technology*** *in* ***Computer Science & Engineering****” is a record of bonafide work carried out by her under my guidance and supervision during the academic year 2020–2021 and it has been found worthy of acceptance according to the requirements of the university.*

**Internal Guide Headof the Department**

**External Examiner**

Annexure – 3

Contents

S.No. Topic Page No.

1. Introduction 1-2
2. System Analysis 3-5
   1. Existing System 4
   2. Proposed System 4
   3. Feasibility Study 5
3. System Requirements Specification 6-7
   1. Software Requirements 7
   2. Hardware Requirements 7
   3. Functional Requirements 7
4. System Design 8-17
   1. Introduction 9
   2. UML Diagrams 10-17
5. System Implementation 18-31
   1. Introduction
   2. Project Modules
   3. Algorithms
6. System Testing32-38
   1. Introduction
   2. Testing Methods
   3. Testing Strategy
7. Conclusion 39-40
8. Bibliography 41-42
9. Appendix 43-47
   1. Introduction to Python
   2. Anaconda
   3. Jupyter Notebook
   4. Matplotlib
   5. CountVectorizer

## INTRODUCTION

### Introduction

The Spam email is one of the significant threats in the world today and has caused tremendous financial losses. Although the methods of confrontation are continually being updated, the results of those methods are not very satisfactory at present. Moreover, spam emails are growing at an alarming rate in recent years. Email Spam has become a major problem nowadays, with rapid growth of internet users, Email spams is also increasing. People are using them for illegal and unethical conducts, phishing and fraud. Sending malicious links through spam emails which can harm our system and can also seek into the system.

Creating a fake profile and email account is much easier for the spammers, they pretend like a genuine person in their spam emails, these spammers target those peoples who are not aware about these frauds. So, it is needed to Identify those spam mails which are fraud, in order to achieve the required target, here we will implement the machine learning algorithms and apply all these algorithms on our data sets and the best algorithm is selected for the email spam detection having best precision and accuracy.

## SYSTEM ANALYSIS

### System Analysis

* 1. **ExistingSystem**

**Gmail Spam Detection System**

**Features :**

1. Gmail has an automatic spam detection system that uses a combination of pattern analysis, user analytics, and virus or phishing detection to send suspicious messages directly to spam.
2. Gmail’s spam detection and filtering, is part of why the email service has reached its current popularity levels.
3. Gmail announced a new feature that will allow users to determine why a message was sent to the spam folder.

**Drawbacks :**

1. Gmail marks valid email messages as spam.
2. Size issues - files need to be small enough to download quickly. Emails containing many images may take too long to load, frustrating your audience and losing their interest.
3. We can only move the incoming spam messages to the spam folder but we can’t prevent receiving spam mails

### ProposedSystem

This project has the aim to achieve spam email classification. In order to do that, a binary classification model is constructed to classify a given mail into spam or non-spam. Since a mail may belongs to twoclasses (spam or non-spam), this is a binary classification problem.

Advantages

To improve the performance of the model.

### FeasibilityStudy

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

* + - OperationalFeasibility
    - EconomicFeasibility
    - TechnicalFeasibility

### Operational Feasibility:

Operational feasibility deals with the study of prospects of the system to be developed. This system operationally predicts the mail whether it is spam or not. Based on the study, the system is proved to be operationallyfeasible.

### Economic Feasibility:

Economic feasibility or Cost-benefit is an assessment of the economic justification for the computer based project. As hardware was installed from the beginning and for lots of purposes thus the cost on hardware is low. So, the project is economically feasible.

### Technical Feasibility:

Technical Feasibility is the process of validating the technology assumptions, architecture and design of a product or project. It was created in the Anaconda. We used jupyter notebook for running the machine learning model. Thus, the project is technically feasible.

## SYSTEM REQUIREMENTS SPECIFICATION

### System Requirements Specification

* 1. **SoftwareRequirements**
     + Anaconda (JupyterNotebook)
     + Language :Python

### HardwareRequirements

* + - Processor :I3 or above
    - RAM :4GB
    - Space on HardDisk:512GB

### FunctionalRequirements

In Software Engineering, a functional requirement defines a function of a software system or its component. A function is described as a set of inputs, behavior, outputs. Functional requirements may be calculations, technical details, data manipulation and supposed to accomplish. The plan for implementing functional requirements is detailed in the system design. In requirements engineering, functional requirement specify particular results of a system. Functional requirements drive the application architecture of asystem.

The following are the functional requirements of our system

* + - User should be able to upload the email.
    - System should able to process theemail.
    - It should be able to classify the email into spam or non-spam.

## SYSTEM DESIGN

### System Design

* 1. **Introduction**

Design is the first step in the development phase of an engineering product or system. Design is the place where quality is considered in the software development. Design is the only way that we can accurately translate user requirements into finished software product or system. Software design serves as the foundation for all the software engineers and software maintenance that steps follow. Without design we risk building an unstable design one that will fail when small changes are made, one that may be difficult to test and one whose quantity cannot be assessed until late in software engineering process.

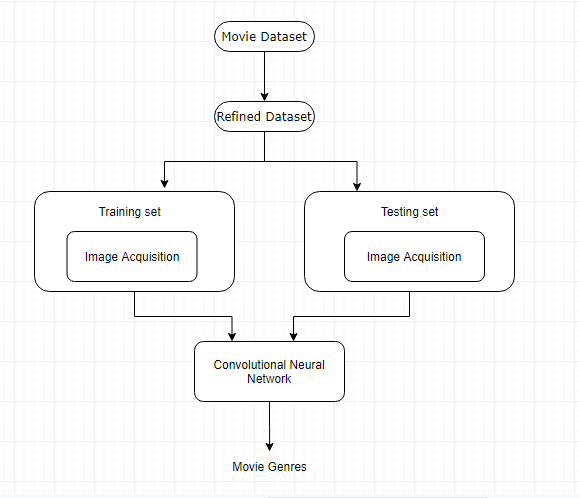


Figure 4.1 system architecture

### Data flow diagrams (UML Diagrams)

**Introduction to UML**

A model is an abstract representation of system, constructed to understand the system priority to building or modifying it. A model is a simplified representation of reality and it provides a means for conceptualization and communication of ideas in a precise and ambiguous form. We build models so that we can better understand the system we are developing. The elements are like components which can be associated in different ways to make a complete UML picture, which is known as diagram. Thus, it is very important to understand the different diagrams to implement the knowledge in real lifesystems.

UML (Unified Modeling Language) is a standard language for specifying, visualizing, constructing, and documenting the artifacts of software systems. It is a method for describing the system architecture in detail using the blueprint. We use UML diagrams to portray the behavior and structure of a system. This is the step while developing any product after analysis. The goal from this is to produce a model of the entities involved in the project which later need to be built. The representation of the entities that are to be used in the product being developed need to be designed.

There are various kinds of methods in software design:

* + - Use caseDiagram
    - ClassDiagram
    - Activity Diagram
    - State ChartDiagram

### Use CaseDiagram

Use Case Diagrams are used to depict the functionality of a system or a part of a system. They are widely used to illustrate the functional requirements of the system and its interaction with external agents(actors).

A use case is basically a diagram representing different scenarios where the system can be used. A use case diagram gives us a high level view of what the system or a part of the system does without going into implementation details. When the initial task is complete, use case diagrams are modelled to present the outsideview.

In brief, the purposes of use case diagrams can be said to be asfollows

* + - * Used to gather the requirements of asystem.
      * Used to get an outside view of asystem.
      * Identify the external and internal factors influencing thesystem.

Use case diagrams commonly contains

* + - * Usecases
      * Actors
      * Dependency, generalization and associationrelationships.

Use cases

A use case is a software and system engineering term that describes how a user uses system to accomplish a particular goal.

Actors

An actor is a person, organization or external system that plays a role in one or more interactions with the system.

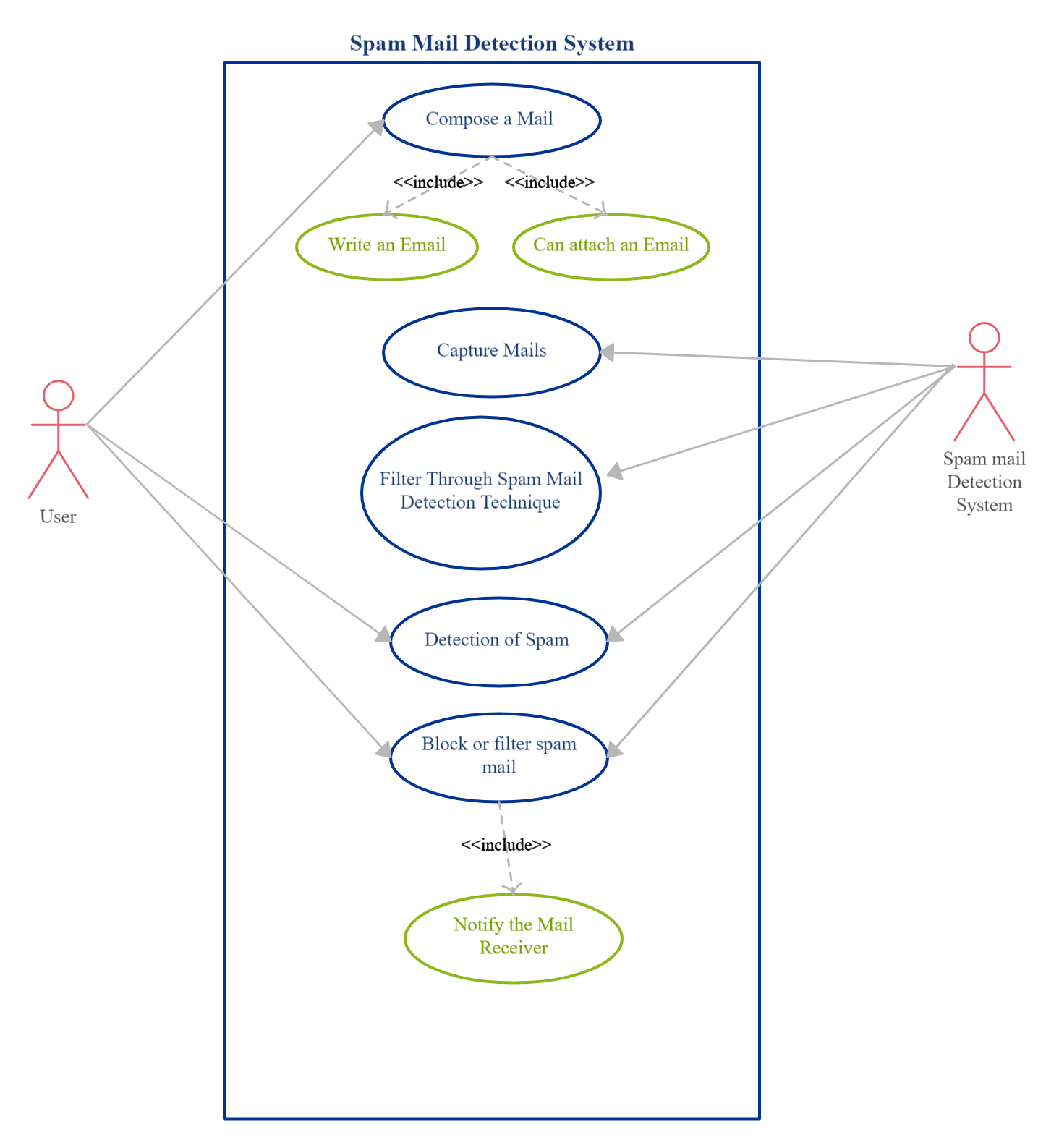


Figure 4.2 Use Case Diagram

* + 1. **ClassDiagram**

Class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing, and documenting different aspects of a system but also for constructing executable code of the software application. Class diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modelling of object-oriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages. It is also known as a structuraldiagram.

Class diagramcontains

* + - * Classes
      * Interfaces
      * Dependency, generalization andassociation.

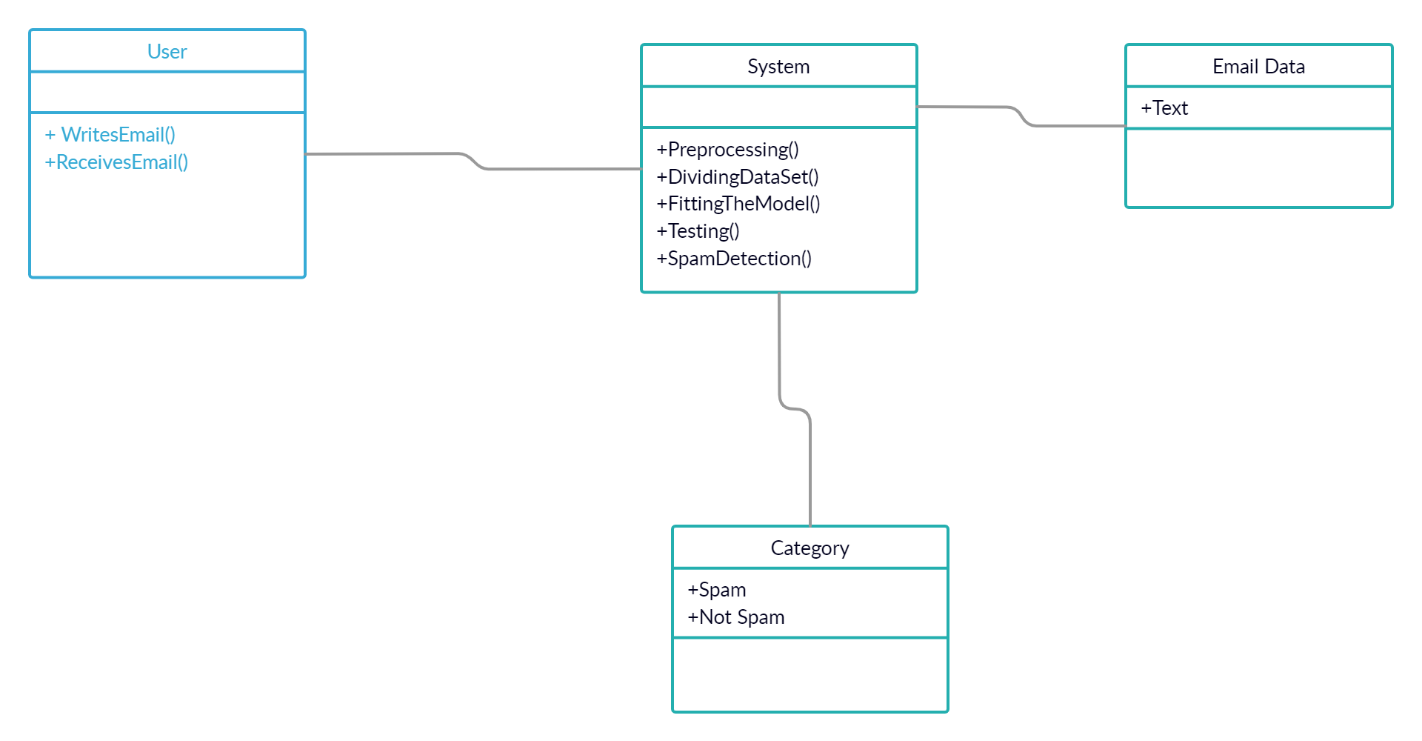


Figure 4.3 Class Diagram

### ActivityDiagram

Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another. This flow can be sequential, branched, or concurrent. Activity diagrams deal with all type of flow control by using different elements such as fork, join, etc. The basic purposes of activity diagram is similar to other four diagrams. It captures the dynamic behavior of the system. Other four diagrams are used to show the message flow from one object to another but activity diagram is used to show message flow from one activity toanother.

Activity is a particular operation of the system. Activity diagrams are not only used for visualizing the dynamic nature of a system, but they are also used to construct the executable system by using forward and reverse engineering techniques. The only missing thing in the activity diagram is the message part. It does not show any message flow from one activity to another. Activity diagram is sometimes considered as the flowchart. Although the diagrams look like a flowchart, they are not. It shows different flows such as parallel, branched, concurrent, and single.

* + - * Describe the sequence from one activity toanother.
      * Draw the activity flow of asystem.
      * Describe the parallel, branched and concurrent flow of thesystem.

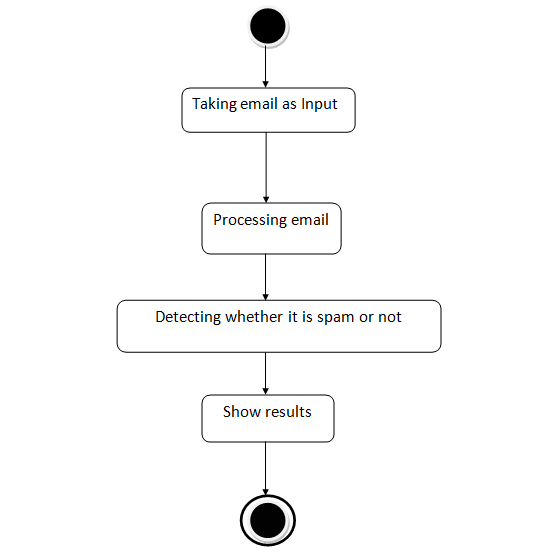


Figure 4.5 Activity Diagram

### State Chart Diagram

State chart diagram describes the flow of control from one state to another state. States are defined as a condition in which an object exists and it changes when some event is triggered. The most important purpose of State chart diagram is to model lifetime of an object from creation to termination. State chart diagrams are also used for forward and reverse engineering of a system. However, the main purpose is to model the reactive system.

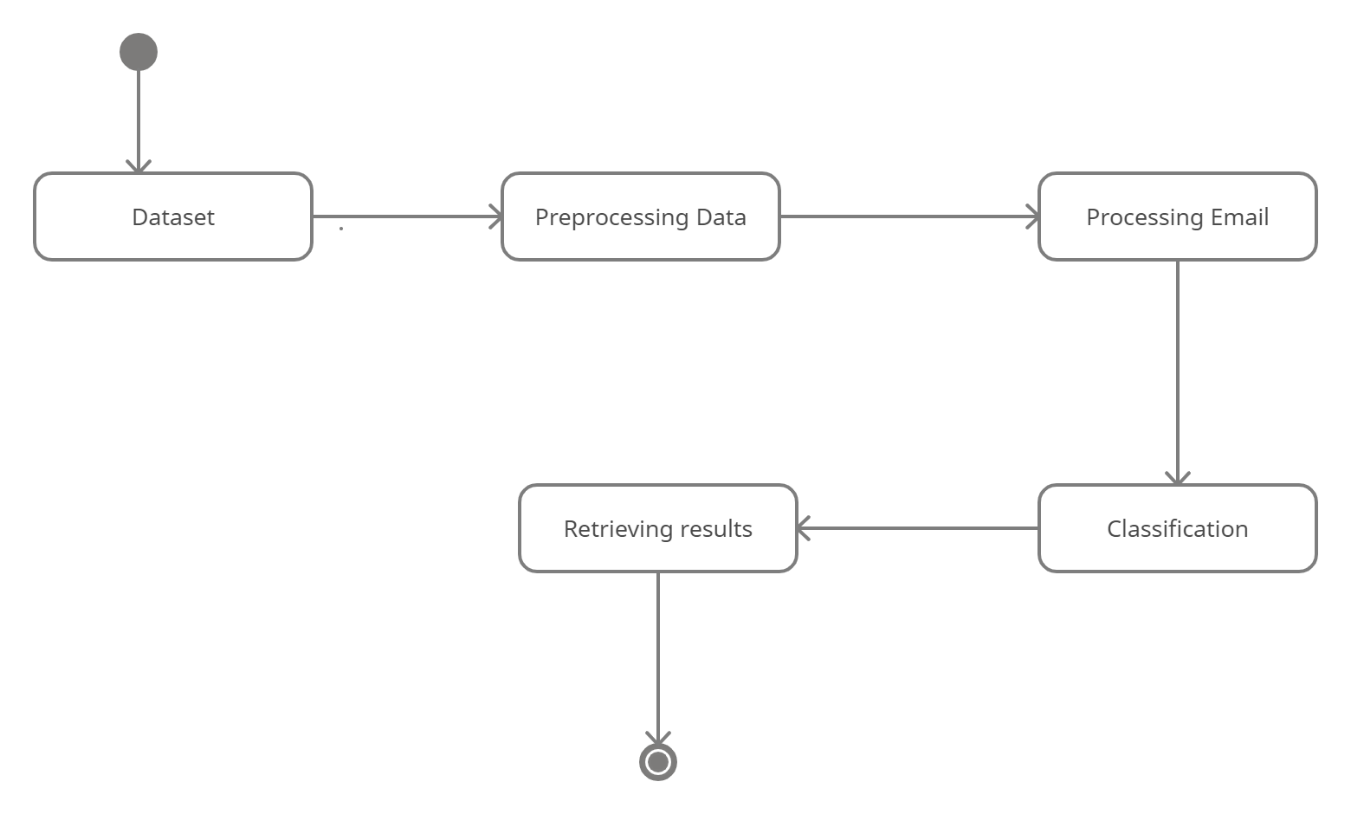


Figure 4.6 State chart diagram

## SYSTEM IMPLEMENTATION

### System Implementation

* 1. **Introduction**

The purpose of system implementation can be summarized as follow making the new system available to the prepared set of users (the deployment), and positioning on- going support and maintenance of the system within the performing organization (the transaction). At a finer necessary to educate the consumer on the use of system, placing the newly developed system into production, confirming that business functions that interact with the system and functioning properly. Transitioning the system support responsibilities involve changing from a system development to the system and maintenance mode of operation, with ownership of the new system moving from the project team to the performingorganization.

A key difference between system implementation and all other phases of lifecycle is that all project activities up to this point have been performed in safe, protected and check your environments. It is through the careful planning, execution and management of system implementation activities that the project team can minimize the likelihood of these occurrences and determine appropriate contingency plans in the event of the problem.

Our project explores if there are certain elements of an email that allow a model to detect the spam email. The input to our algorithm is an email text andour model outputs classification of email. We apply various binary classification algorithms and strategies to learn features of the mail to make our predictions.

### ProjectModules

1.Data Gathering. 2.Data Preprocessing.

3.Dividing dataset into training data and testing data. 4.Designing the model.

5.Training and testing the model. 6.Predicting the spam mail.

### DataGathering

Data gathering generally involves collection of data from various sources. In this project we are gathering a dataset which consists of data about movies. A dataset (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. Data sets can also consist of a collection of documents orfiles.

The dataset collected for this project consists of over 5,572 instances of emails.



Figure 5.1 Instance of a dataset

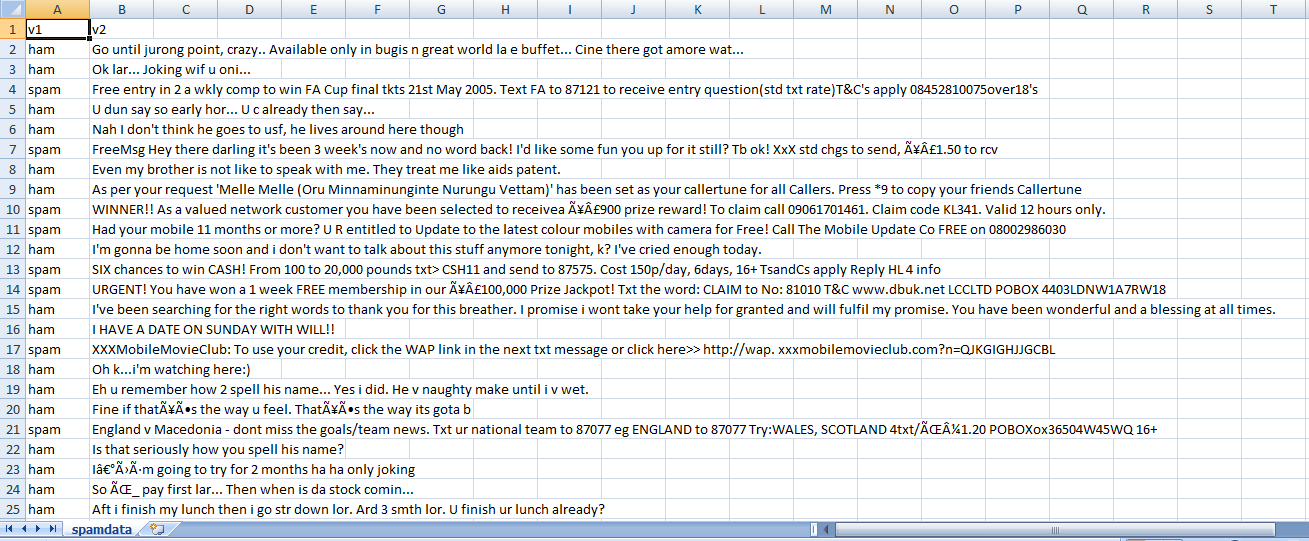


Figure 5.2 Instance of emails

### Data preprocessing

Data preprocessing is a technique that is used to convert raw data into a clean dataset. The data is gathered from different sources is in raw format which is not feasible for the analysis. Pre-processing for this approach takes 4 simple yet effective steps.

### Attribute selection

Some of the attributes in the initial dataset that were not pertinent (relevant) to the experiment goal were ignored.

### Cleaning missing values:

In some cases, the dataset contains missing values. We need to be equipped to handle the problem when we come across them. Obviously, you could remove the entire line of data but what if you're inadvertently removing crucial information after all we might not need to try to do that. One in every of the foremost common plan to handle the matter is to require a mean of all the values of the same column and have it to replace the missing data. The library used for the task is called scikit Learn preprocessing. It contains a class called Imputer which will help us take care of the missingdata.

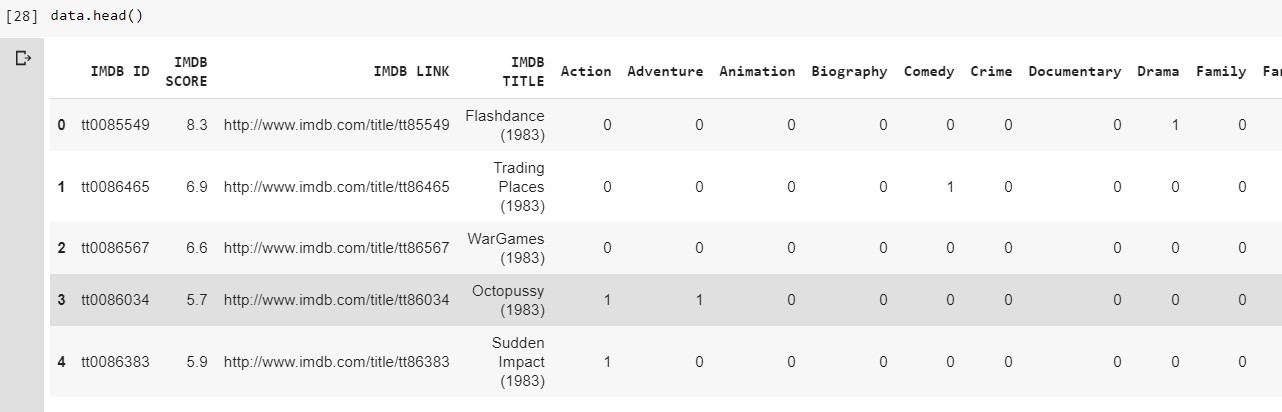


Figure 5.3 Preprocessed dataset

### Dividing dataset into Training and Test data

Splitting the Dataset into Training set and Test set We will train our machine learning models on our training set our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to examine how accurately it will predict. A general rule of the thumb is to assign 70% of the dataset to training set and therefore the remaining 30% to test set.

**Training set** — a subset to train a model.

**Test set** — a subset to test the trained model.

### Feature Extraction

The process of feature extraction is useful when you need to reduce the number of resources needed for processing without losing important or relevant information. Feature extraction can also reduce the amount of redundant data for a given analysis. Also, the reduction of the data and the machine’s efforts in building variable combinations (features) facilitate the speed of learning and generalization steps in the machine learning process.

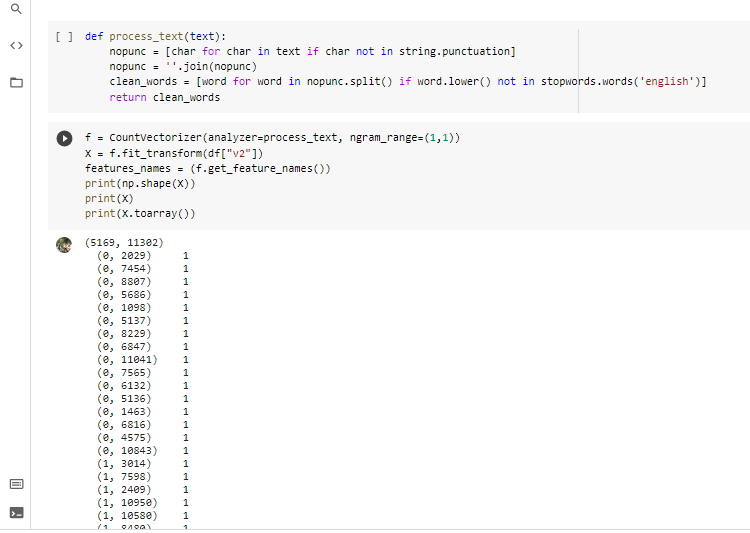


Figure 5.4 Feature Extraction

### Training and testing the model

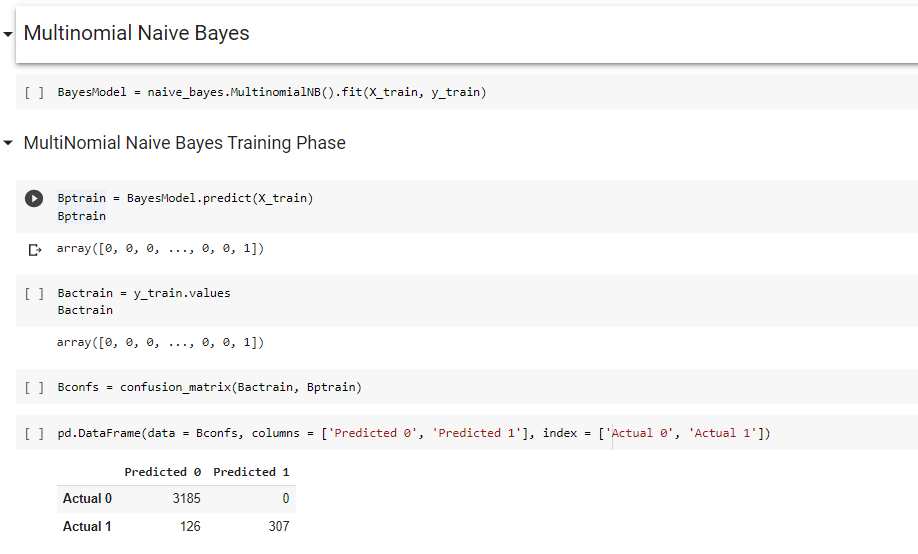
****

Figure 5.6 MNB Training phase

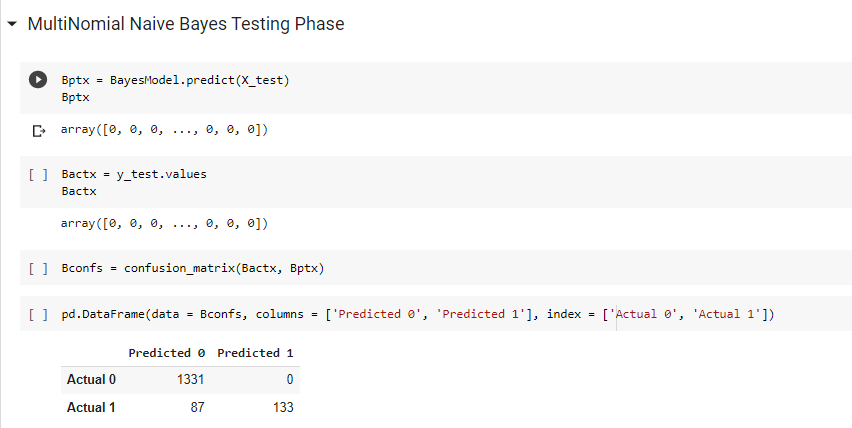


Figure 5.7 MNB Testing phase



Figure 5.8KNN Training phase

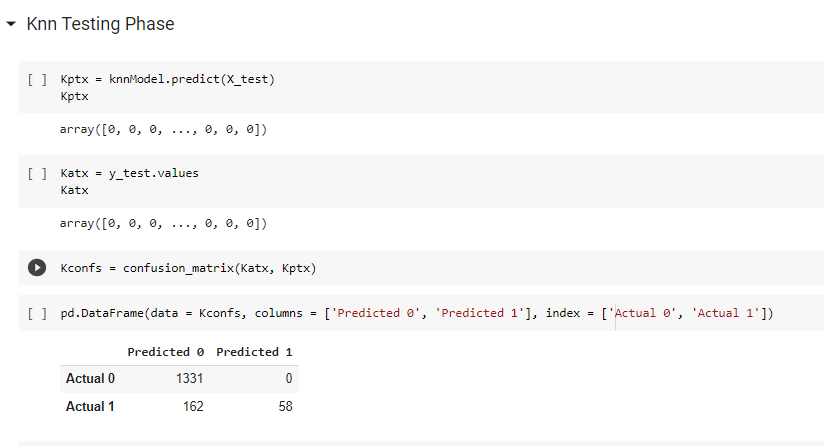


Figure 5.9KNNTesting phase



Figure 5.10DecisionTree with gini index Training phase

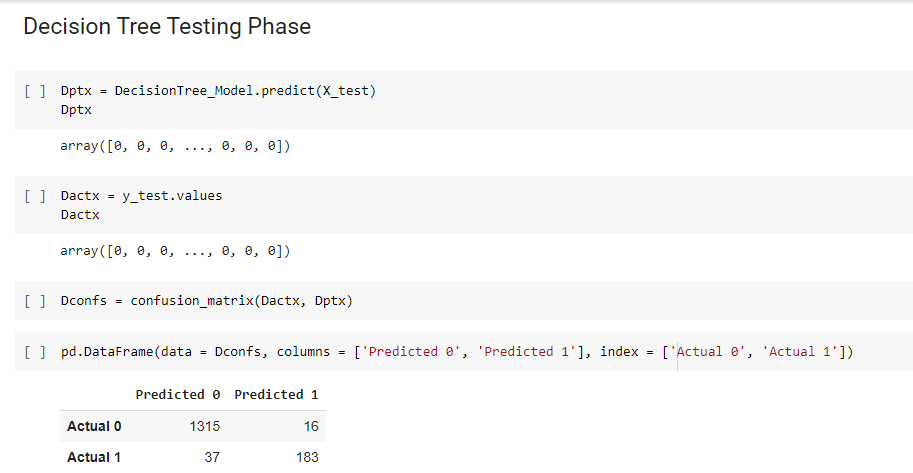


Figure 5.11DecisionTree with gini Testing phase



Figure 5.12DecisionTree (entropy) Training phase

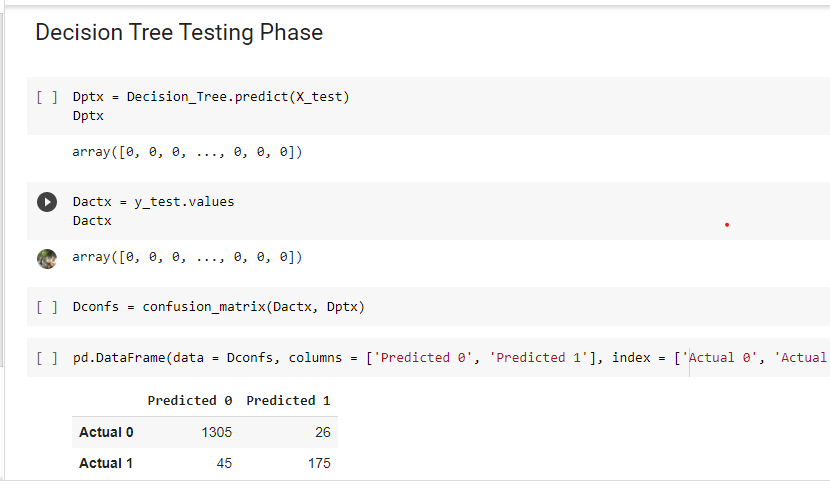


Figure 5.13DecisionTree(entropy) Testing phase



Figure 5.14SVC Training phase

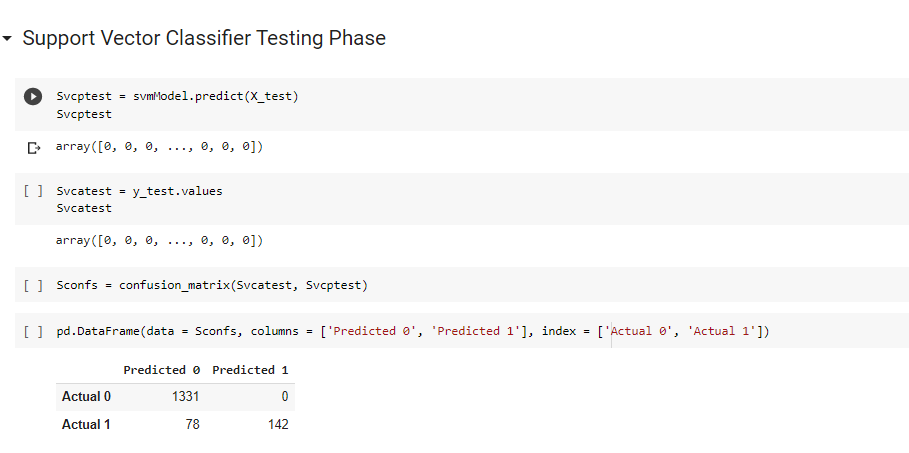


Figure 5.15SVCTesting phase

### Algorithms

**Multinomial Naïve Bayes Algorithm**

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature.

### DecisionTree Algorithm

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

A tree is composed of nodes, and those nodes are chosen looking for the optimum split of features. For that purpose, different criteria exist. In the decision tree Python implementation of the scikit-learn library, this is made by the parameter ‘creation’. This parameter is the function used to measure the quality of a split and it allows users to choose between gini or entropy.

### Support Vector Machine Algorithm

Support Vector Machine is a supervised machine learning algorithm, which can be used for both classification or regression challenges. However,  it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of

a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

### K-nearest neighbors Algorithm

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

## TESTING

### Testing

* 1. **Introduction**

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring thatthe

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testingrequirement.

### TESTING METHODOLOGIES

The following are the Testing Methodologies:

* + - UnitTesting.
    - IntegrationTesting.
    - User AcceptanceTesting.
    - OutputTesting.
    - ValidationTesting.

### Unit Testing

Unit testing focuses verification effort on the smallest unit of Software design that is the module. Unit testing exercises specific paths in a module’s control structure to ensure complete coverage and maximum error detection. This test focuses on each module individually, ensuring that it functions properly as a unit. Hence, the naming is Unit Testing.

During this testing, each module is tested individually and the module interfaces are verified for the consistency with design specification. All important processing path are tested for the expected results. All error handling paths are also tested.

### Integration Testing

Integration testing addresses the issues associated with the dual problems of verification and program construction. After the software has been integrated a set of high order tests are conducted. The main objective in this testing process is to take unit tested modules and builds a program structure that has been dictated by design.

#### The following are the types of Integration Testing:

1. **Top DownIntegration**

This method is an incremental approach to the construction of program structure. Modules are integrated by moving downward through the control hierarchy, beginning with the main program module. The module subordinates to the main program module are incorporated into the structure in either a depth first or breadth first manner.

In this method, the software is tested from main module and individual stubs are replaced when the test proceeds downwards.

#### Bottom-upIntegration

This method begins the construction and testing with the modules at the lowest level in the program structure. Since the modules are integrated from the bottom up, processing required for modules subordinate to a given level is always available and the

need for stubs is eliminated. The bottom up integration strategy may be implemented with the following steps:

* + The low-level modules are combined into clusters into clusters that perform a specific Softwaresub-function.
  + A driver (i.e.) the control program for testing is written to coordinate test case input andoutput.
  + The cluster istested.
  + Drivers are removed and clusters are combined moving upward in the program structure

The bottom up approaches test each module individually and then each module is module is integrated with a main module and tested forfunctionality.

#### OTHER TESTING METHODOLOGIES

**User Acceptance Testing**

User Acceptance of a system is the key factor for the success of any system. The system under consideration is tested for user acceptance by constantly keeping in touch with the prospective system users at the time of developing and making changes wherever required. The system developed provides a friendly user interface that can easily be understood even by a person who is new to the system.

#### Output Testing

After performing the validation testing, the next step is output testing of the proposed system, since no system could be useful if it does not produce the required output in the specified format. Asking the users about the format required by them tests the outputs generated or displayed by the system under consideration. Hence the output format is considered in 2 ways – one is on screen and another in printed format.

#### Validation Checking

Validation checks are performed on the following fields.

#### Text Field:

The text field can contain only the number of characters lesser than or equal to its size. The text fields are alphanumeric in some tables and alphabetic in other tables. Incorrect entry always flashes and error message.

#### Numeric Field:

The numeric field can contain only numbers from 0 to 9. An entry of any character flashes an error messages.

The individual modules are checked for accuracy and what it has to perform. Each module is subjected to test run along with sample data. The individually tested modules are integrated into a single system. Testing involves executing the real data information is used in the program the existence of any program defect is inferred from the output. The testing should be planned so that all the requirements are individuallytested.

A successful test is one that gives out the defects for the inappropriate data and produces and output revealing the errors in the system.

#### Preparation of Test Data

Taking various kinds of test data does the above testing. Preparation of test data plays a vital role in the system testing. After preparing the test data the system under study is tested using that test data. While testing the system by using test data errors are again uncovered and corrected by using above testing steps and corrections are also noted for future use.

#### Using Live Test Data:

Live test data are those that are actually extracted from organization files. After a system is partially constructed, programmers or analysts often ask users to key in a set of data from their normal activities. Then, the systems person uses this data as a way to partially test the system. In other instances, programmers or analysts extract a set of live data from the files and have them enteredthemselves.

It is difficult to obtain live data in sufficient amounts to conduct extensive testing. And, although it is realistic data that will show how the system will perform for the typical processing requirement, assuming that the live data entered are in fact typical, such data generally will not test all combinations or formats that can enter the system. This bias toward typical values then does not provide a true systems test and in fact ignores the cases most likely to cause systemfailure.

#### Using Artificial Test Data:

Artificial test data are created solely for test purposes, since they can be generated to test all combinations of formats and values. In other words, the artificial data, which can quickly be prepared by a data generating utility program in the information systems department, make possible the testing of all login and control paths through the

program. The most effective test programs use artificial test data generated by persons other than those who wrote the programs. Often, an independent team of testers formulates a testing plan, using the systems specifications. The package “Virtual Private Network” has satisfied all the requirements specified as per software requirement specification and was accepted.

#### USER TRAINING

Whenever a new system is developed, user training is required to educate them about the working of the system so that it can be put to efficient use by those for whom the system has been primarily designed. For this purpose the normal working of the project was demonstrated to the prospective users. Its working is easily understandable and since the expected users are people who have good knowledge of computers, the use of this system is very easy.

#### MAINTAINENCE

This covers a wide range of activities including correcting code and design errors. To reduce the need for maintenance in the long run, we have more accurately defined the user’s requirements during the process of system development. Depending on the requirements, this system has been developed to satisfy the needs to the largest possible extent. With development in technology, it may be possible to add many more features based on the requirements in future. The coding and designing is simple and easy to understand which will make maintenanceeasier.

#### TESTING STRATEGY:

A strategy for system testing integrates system test cases and design techniques into a well planned series of steps that results in the successful construction of software. The testing strategy must co-operate test planning, test case design, test execution, and the resultant data collection and evaluation .A strategy for software testing must accommodate low-level tests that are necessary to verify that a small source code segment has been correctly implemented as well as high level tests that validate major system functions against userrequirements.

Software testing is a critical element of software quality assurance and represents the ultimate review of specification design and coding. Testing represents an interesting anomaly for the software. Thus, a series of testing are performed for the proposed system before the system is ready for user acceptance testing.

#### SYSTEM TESTING:

Software once validated must be combined with other system elements (e.g. Hardware, people, database). System testing verifies that all the elements are proper and that overall system function performance is achieved. It also tests to find discrepancies between the system and its original objective, current specifications and system documentation.

#### UNIT TESTING:

In unit testing different are modules are tested against the specifications produced during the design for the modules. Unit testing is essential for verification of the code produced during the coding phase, and hence the goals to test the internal logic of the modules. Using the detailed design description as a guide, important Conrail paths are tested to uncover errors within the boundary of the modules. This testing is carried out during the programming stage itself. In this type of testing step, each module was found to be working satisfactorily as regards to the expected output from themodule.

In Due Course, latest technology advancements will be taken into consideration. As part of technical build-up many components of the networking system will be generic in nature so that future projects can either use or interact with this. The future holds a lot to offer to the development and refinement of thisproject.

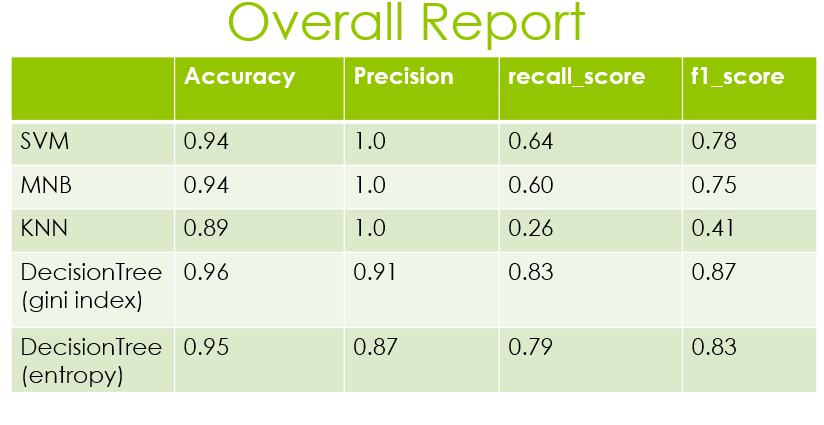
## CONCLUSION

### CONCLUSION

A study on detection of spam emails was modeled as a binary classification task. The experiment was conducted on a dataset containing 5572 emails, classified into either spam or non-spam. We have done feature extraction on the weights of most common words in both spam emails and non-spam emails separately and done the feature transformation in order to fit the model.

In particular, our models did well on the classification. This indicates the potential for improved performance given more computational resources and a more balanced dataset.

On the basis of all the classification reports generated by the classification algorithms like MNB, SVM, KNN, DecisionTree algorithms, we got the accuracy values as follows :



## BIBILOGRAPHY

### Bibilography

1. <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>
2. <https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623>
3. <https://towardsdatascience.com/feature-extraction-techniques-d619b56e31be>
4. <https://youtu.be/exHwwy9kVcg>
5. <https://youtu.be/RZYjsw6P4nI>

## APPENDIX

### APPENDIX

**9.1 Introduction to python**

Python is an open source, high-level programming language developed by Guido van Rossum in the late 1980s and presently administered by Python Software Foundation. It came from the ABC language that he helped create early on in his career. Python is a powerful language that you can use to create games, write GUIs, and develop web applications.

It is a high-level language. Reading and writing codes in Python is much like reading and writing regular English statements. Because they are not written in machine- readable language, Python programs need to be processed before machines can run them. Python is an interpreted language. This means that every time a program is run, its interpreter runs through the code and translates it into machine readable bytecode.

Python is an object-oriented language that allows users to manage and control data structures or objects to create and run programs. Everything in Python is latest version of Python in fact, first class. All objects, data types, functions, methods, and classes take equal position in Python. Programming languages are created to satisfy the needs of programmers and users for an effective tool to develop applications that impact lives, lifestyles, economy, and society. They help make lives better by increasing productivity, enhancing communication, and improving efficiency. Languages die and become obsolete when they fail to live up to expectations and are replaced and superseded by languages that are morepowerful.

Python is a programming language that has stood the test of time and has remained relevant across industries and businesses and among programmers, and individual users. It is a living, thriving, and highly useful language that is highly recommended as a first programming language for those who want todive

In to and experience programming. Advantages of Using Python Here are reasons why you would prefer to learn and use Python over other high-level languages

#### Readability

Python programs use clear, simple, and concise instructions that are easy to read even by those who have no substantial programming background. Programs written in Python are, therefore, easier to maintain, debug, orenhance.

#### Higher productivity

Codes used in Python are considerably shorter, simpler, and less verbose than other high level programming languages such as Java and C++. In addition, it has well-designed built-in features and standard library as well as access to third party modules and source libraries. These features make programming in Python more efficient.

#### Less learning time

Python is relatively easy to learn. Many find Python a good first language for learning programming because it uses simple syntax and shorter codes. Python works on Windows, Linux/UNIX, Mac OS X, other operating systems and small form devices. It also runs on microcontrollers used in appliances, toys, remote controls, embedded devices, and other similardevices.

#### Installing Python in Windows

To install Python, you must first download the installation package of your preferred version from this link: http[s://w](http://www.python.org/downloads/)ww.p[ython.org/downloads/](http://www.python.org/downloads/) On this page, you will be asked to choose between the two latest versions for Python 2 and 3: Python 3.5.1 and Python 2.7.11. Alternatively, if you are looking for a specific release, you can scroll down the page to find download links for earlier versions. You would normally opt to download the latest version, which is Python 3.5.1. This was released on December 7, 2015. However, you may opt for the latest version of Python 2, 2.7.11. Your preferences will usually depend on which version will be most usable for your project. While Python 3 is the present and future of the language, issues such as third-party utility or compatibility may require you to download Python 2.

### 9.2 Anaconda:

Anaconda is a free open source distribution of the python and R programming languages for large-scale data processing, predictive analysis, and scientific computing that aims to simplify package management and deployment. Package versions are managed by the package management system. Where packages, notebooks, projects and environments are shared. Powerful collaboration and package management for open source and privateprojects.

### 9.3 Jupyter Notebook:

The Jupyter Notebook is an open source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. We can perform data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning and much more. It is available in anaconda software and can be usedeasily.

### Getting Started:

There are two ways that python is commonly used. The first is an interactive command environment, each as Python or IDLE, which are commonly bundled with the Python interpreter. Starting Python with one of these (using Start/Python in Windows, or by typing python at a command prompt (which will be shown as >>>). Unlike with C or Java, you can type commands at this prompt and the interpreter will run the commands and display the results, if any, on the screen. You can write functions in a text editor and run them from the command prompt by calling them by name.

### 9.4 Matplotlib:

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension numpy.

%matplotlib inline: This turns on inline plotting, where plot graphics will appear in your notebook. This has important implications for interactivity. For inline plotting, commands cells below the cell that outputs a plot will not affect the plot.

For example, changing the map color map is not possible from cells below the cells below the cell that creates aplot.

Pyplot is a module in the matplotlib package. That’s why you often see matplotlib.pyplot in code. The module provides an interface that allows you to implicitly and automatically create figures axes a achieve the desired plot. This is especially handy when you want to quickly plot something without instantiating any figures or axes as you saw in the example in the first section of this tutorial. You see, you haven’t explicitly specified these components, yet you manage to output a plot that you have even customized! The defaults are initialized and any customizations that you do, will be done with the current Figures and Axes is mind.

Lastly, Pylab is another module, but it gets installed alongside the matplotlib package. It bulks imports pyplot and the numpy library and was generally recommended when you were working with arrays, doing mathematics interactively and wanted across to plotting features.

### 9.5 CountVectorizer:

Scikit-learn’s CountVectorizer is used to convert a collection of text documents to a vector of term/token counts. It also enables the ​pre-processing of text data prior to generating the vector representation. This functionality makes it a highly flexible feature representation module for text.