

# **EMAIL PHISHING MESSAGES CLASSIFICATION USING MACHINE LEARNING**

Submitted in partial fulfillment of the requirements for the award of  
Bachelor of Engineering degree in Computer Science and Engineering

By

**NARREDDY MURALI KRISHNA REDDY (39110680)  
MADALLAPALLI SUSHANTH (39110997)**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
SCHOOL OF COMPUTING**

## **SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY  
(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE  
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,  
CHENNAI – 600119  
APRIL - 2023**



**SATHYABAMA**  
INSTITUTE OF SCIENCE AND TECHNOLOGY  
(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE

[www.sathyabama.ac.in](http://www.sathyabama.ac.in)

---

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonafide work of **Narreddy Murali Krishna Reddy (39110680) and Madallapalli Sushanth (39110997)** who carried out the Project Phase-2 entitled **"Email Phishing Messages Classification using Machine Learning "** under my supervision from **January 2023 to April 2023.**

**Internal Guide**

**Ms.C.A.Daphine Desona Clemency, M.E**

**Head of the Department**

**Dr. L. LAKSHMANAN, M.E., Ph.D.**



---

**Submitted for Viva voce Examination held on 20.4.2023**

**Internal Examiner**

**External Examiner**

## DECLARATION

I, **Narreddy Murali Krishna Reddy (39110680)**, hereby declare that the Project Phase-2 Report entitled “**Email Phishing Messages Classification using Machine Learning**” done by me under the guidance of **Ms.C.A.Daphine Desona Clemency M.E.** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

**DATE:20.4.2023**

**PLACE: Chennai**

A handwritten signature in black ink, reading "Madallapalli Sushanth". The signature is written in a cursive style with a horizontal line underneath the name.

**SIGNATURE OF THE CANDIDATE**

## ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph. D, Dean**, School of Computing, **Dr. L. Lakshmanan M.E., Ph.D.**, Head of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Ms.C.A.Daphine Desona Clemency M.E.** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my phase-2 project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

## ABSTARCT

Online review systems play an important role in affecting consumers behaviors and decision making, attracting many spammers to insert fake reviews to manipulate review content and ratings. To increase utility and improve user experience, some online review systems allow users to form social relationships between each other and encourage their interactions. In this paper, we aim at providing an efficient and effective method to identify review spammers by incorporating social relations based on two assumptions that people are more likely to consider reviews from those connected with them as trustworthy, and review spammers are less likely to maintain a large relationship network with normal users.

Vehicle reconciliation strategies have been involved a few times in spam channels to incorporate approaching/active messages, for example, spam and spam bunches. This technique expresses that each bunch contains little miniature groups, and each miniature group is dispersed. Notwithstanding, this thought ought not be trifled with, and the miniature group might have a lopsided dispersion. To build the respectability of the main strategy for appropriating the Internet class, we suggest supplanting the Euclidean space with a succession of models that incorporate into the miniature bunch connected with the circulation. Here, the Naïve Bayes classification has been carried out to carry out miniature bunches across the line. While these INBs can decide the distance and limits of micro clusters, Euclidean space considers the overall worth of the group and misdirects the bigger micro cluster. In this report, Den Stream is upheld by a committed framework called INB Den Stream. To represent the presentation of INB-Den Stream, current techniques, their exhibition has been estimated as far as quality, honesty as a general rule, memory as a rule, F1 measurements., markers, and complex estimations. The relatively close outcomes show that our techniques outflank its rivals in the set figures

## TABLE OF CONTENTS

Chapter No	Title	Page No
	<b>ABSTRACT</b>	v
	<b>LIST OF ABBREVIATIONS</b>	viii
	<b>LIST OF FIGURES</b>	ix
1	<b>INTRODUCTION</b>	1
	1.1 What is can social network be used for ?	3
	1.2 Secure Computing	4
	1.3 some basic needs for secure computing	4
	1.4 Key steps to spam mail detection	6
2	<b>LITERATURE SURVEY</b>	
	2.1 Inferences from literature survey	8
	2.2 Open problems in Existing system	10
3	<b>REQUIREMENTS ANALYSIS</b>	
	3.1 Software Requirements Specification Documents	12
	3.1.1 Python	12
	3.1.2 Why is python	13
	3.1.3 Advantages of python	13
	3.1.4 Disadvantages of python	16
	3.1.5 Install python step-by-step in windows and mac	16
	3.1.6 How to install python on windows and mac	17
	3.1.7 Download the correct version into windows and mac	17
	3.1.8 Pycharm Ide	19
	3.1.9 Installing pycharm ide	20
	3.1.10 Machine learning libraries	23
	3.2 System use case	25
4	<b>DESCRIPTION OF PROPOSED SYSTEM</b>	
	4.1 selected methodology or process model	27
	4.2 Architecture of proposed model	28
	4.3 Description of software for Implementation and Testing Plan of the Proposed System	28

	4.4 Project Management plan	32
	4.5 Financial report on estimated costing	33
5	<b>IMPLEMENTATION DETAILS</b>	
	5.1 Development and Deployment Setup	34
	5.1 Algorithms	35
	5.1.1 Random forest algorithm	35
	5.1.2 Why random forest	35
	5.1.3 Advantages of random forest	36
	5.1.4 Disadvantages of random forest	36
	5.2 Decision Tree classification Algorithm	37
	5.2.1 Why use Decision Trees?	37
	5.2.2 Advantages of Decision tree	38
	5.2.3 Disadvantages of Decision tree	38
	5.3 Testing	39
6	<b>RESULTS AND DISCUSSIONS DETAILS</b>	41
7	<b>CONCLUSION</b>	
	7.1 Conclusion	42
	7.2 Future Work	42
	7.3 Research issues	43
	7.4 Implementation issues	44
	<b>REFERENCES</b>	45
	<b>APPENDIX</b>	
	<b>A.SOURCE CODE</b>	47
	<b>B.SCREENSHOTS</b>	50
	<b>C.RESEARCH PAPER</b>	53

## LIST OF ABBREVIATIONS

Abbreviation No	Abbreviation
1	DTC – Decision tree classifier
2	KNN – K-Nearest neighbor
3	LTSM – Long term short memory
4	NLP – Neural network Processing
5	RF – Random Forest



## LIST OF FIGURES

Figure No	Figure name	Page No
1.1	Email spam modeling	3
1.2	Supervised Learning	6
1.3	Unsupervised Learning	7
3.1	python	12
3.2	Python Official link	17
3.3	Python version Download	18
3.4	Pycharm Official Link	20
3.5	Pycharm project Setup	20
3.6	ML modules	23
3.7	NLP	24
4.1	Architecture	28
4.2	Local host web server	31
4.3	Local host output console	32
5.1	Random forest	37
5.2	Decision Tree	38

## **CHAPTER-1**

### **INTRODUCTION**

A social network service as a service which focuses on the building and verifying of online social networks for communities of people who share interests and activities, or who are interested in exploring the interests and activities of others, and which necessitates the use of software.

A report published by OCLC provides the following definition of social networking sites: Web sites primarily designed to facilitate interaction between users who share interests, attitudes and activities, such as Facebook, Maxi and Myspace.

Email or electronic mail spam refers to the “using of email to send unsolicited emails or advertising emails to a group of recipients. Unsolicited emails mean the recipient has not granted permission for receiving those emails. “The popularity of using spam emails is increasing since last decade. Spam has become a big misfortune on the internet. Spam is a waste of storage, time and message speed.

Automatic email filtering may be the most effective method of detecting spam but nowadays spammers can easily bypass all these spam filtering applications easily. Several years ago, most of the spam can be blocked manually coming from certain email addresses. Machine learning approach will be used for spam detection. Major approaches adopted closer to junk mail filtering encompass “text analysis, white and blacklists of domain names, and community-primarily based techniques”.

Text assessment of contents of mails is an extensively used method to the spams. Many answers deployable on server and purchaser aspects are available.

However, rejecting sends essentially dependent on content examination can be a difficult issue in the event of bogus positives. Regularly clients and organizations would not need any legitimate messages to be lost. The boycott approach has been probably the soonest technique pursued for the separating of spams.

The technique is to acknowledge all the sends other than those from the area/electronic mail ids. Expressly boycotted. With more up to date areas coming into the classification of spamming space names this technique keeps an eye on no longer work so well.

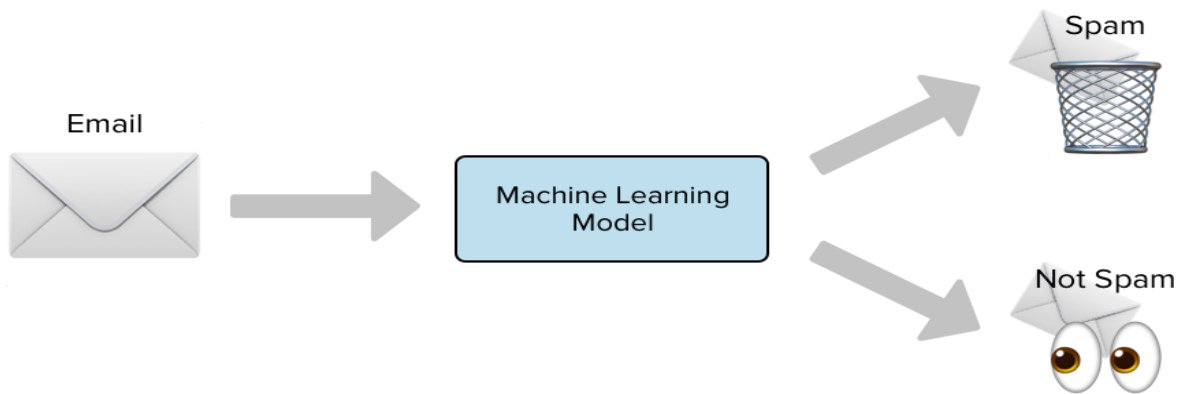
The white list approach is the approach of accepting the mails from the domain names/addresses openly whitelisted and place others in a much less importance queue, that is delivered most effectively after the sender responds to an affirmation request sent through the “junk mail filtering system”

Spam and Ham: According to Wikipedia “the use of electronic messaging systems to send unsolicited bulk messages, especially mass advertisement, malicious links etc.” are called as spam. “Unsolicited means that those things which you didn’t asked for messages from the sources. So, if you do not know about the sender the mail can be spam. People generally don’t realize they just signed in for those mailers when they download any free services, software or while updating the software.

As the Internet continues to grow in both size and importance, the quantity and impact of online reviews continually increases.

Reviews can influence people across a broad spectrum of industries, but are particularly important in the realm of e-commerce, where comments and reviews regarding products and services are often the most convenient, if not the only, way for a buyer to make a decision on whether or not to buy them.

Online reviews may be generated for a variety of reasons. Often, in an effort to improve and enhance their businesses, online retailers and service providers may ask their customers to provide feedback about their experience with the products or services they have bought, and whether they were satisfied or not. Customers may also feel inclined to review a product or service if they had an exceptionally good or bad experience with it.



**Fig 1.1 Email Spam Modeling**

### **1.1 What Can Social Networks Be Used For ?**

Social networks can provide a range of benefits to members of an organization :-

***Support for members of an organization:*** Social networks can potentially be used by all members of an organization, and not just those involved in working with students. Social networks can help the development of communities of practice.

***Engaging with others:*** Passive use of social networks can provide valuable business intelligence and feedback on institutional services.

***Ease of access to information and applications:*** The ease of use of many social networking services can provide benefits to users by simplifying access to other tools and applications. The Facebook Platform provides an example of how a social networking service can be used as an environment for other tools.

***Common interface:*** A possible benefit of social networks may be the common interface which spans work / social boundaries. Since such services are often used in a personal capacity the interface and the way the service works may be familiar, thus minimizing training and support needed to exploit the services in a professional context. This can, however, also be a barrier to those who wish to have strict boundaries.

## 1.2 Secure Computing:-

Computer security is information security as applied to computers and networks. The field covers all the processes and mechanisms by which computer-based equipment, information and services are protected from unintended or unauthorized access, change or destruction. Computer security also includes protection from unplanned events and natural disasters. Otherwise, in the computer industry, the term security or the phrase computer security -- refers to techniques for ensuring that data stored in a computer cannot be read or compromised by any individuals without authorization. Most computer security measures involve data encryption and passwords. Data encryption is the translation of data into a form that is unintelligible without a deciphering mechanism. A password is a secret word or phrase that gives a user access to a particular program or system.

## 1.3 Some Basic needs for security computing:-

**Physical Security:-** Technical measures like login passwords, anti-virus are essential. However, a secure physical space is the first and more important line of defense. Is the place you keep your workplace computer secure enough to prevent theft or access to it while you are away. While the Security Department provides coverage across the Medical center, it only takes seconds to steal a computer, particularly a portable device like a laptop. A computer should be secured like any other valuable possession when you are not present.

**Access Passwords:-** The networks and shared information systems are protected in part by login credentials user-IDs and passwords. Access passwords are also an essential protection for personal computers in most circumstances. Offices are usually open and shared spaces, so physical access to computers cannot be completely controlled.

**Anti-Virus Software:-** Up-to-date, properly configured anti-virus software is essential. While we have server-side anti-virus software on our network computers, you still need it on the client side .

**Firewalls:-** Anti-virus products inspect files on your computer and in email. Firewall software and hardware monitor communications between your computer and the outside world. That is essential for any networked computer.

**Software Updates:-** It is critical to keep software up to date, especially the operating system, anti-virus and anti-spyware, email and browser software. The newest versions will contain fixes for discovered vulnerabilities. Almost all anti-virus have automatic update features,. Keeping the digital patterns of malicious software detectors up-to-date is essential for these products to be effective.

**Keep Secure Backup's:-** Even if you take all these security steps, bad things can still happen. Be prepared for the worst by making backup copies of critical data, and keeping those backup copies in a separate, secure location. For example, use supplemental hard drives to store critical, hard-to-replace data.

**Prying Eye Protection:-** Because we deal with all facets of clinical, research, educational and administrative data here on the medical campus, it is important to do everything possible to minimize exposure of data to unauthorized individuals.

**Enabling the safe operation of applications:-** The organization is under immense pressure to acquire and operates integrated, efficient and capable applications. The modern organization needs to create an environment that safeguards application using the organizations IT systems, particularly those application that serves as important elements of the infrastructure of the organization.

**Protecting the data that the organization collect and use:-** Data in the organization can be in two forms are either in rest or in motion, the motion of data signifies that data is currently used or processed by the system. The values of the data motivated the attackers to steal or corrupts the data. This is essential for the integrity and the values of the organization's data. Information security ensures the protection of both data in motion as well as data in rest.

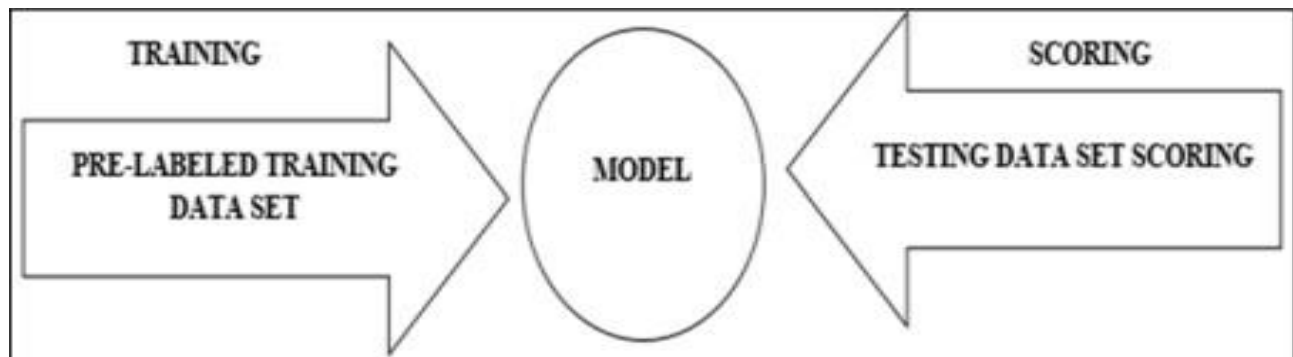
## 1.4 Key steps to Spam Mail Detection:-

**Email Filtering:-** One of the primary methods for spam mail detection is email filtering. It involves categorize incoming emails into spam and non-spam. Machine learning algorithms can be trained to filter out spam mails based on their content and metadata.

**Text Classification:-** Text classification is a supervised learning technique used for spam detection. It involves labelling emails as spam or non-spam based on their features, such as the presence of certain keywords, tone, or grammar.

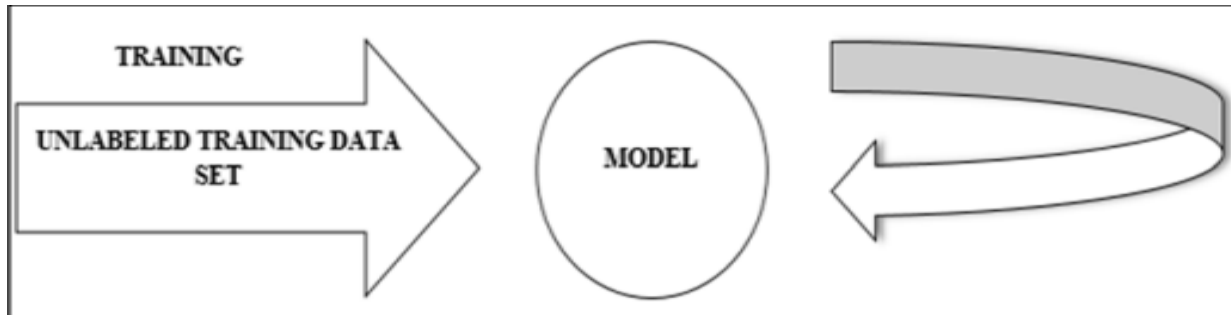
**Feature Engineering:-** Feature engineering is the process of selecting relevant features from the email to classify it as spam or non-spam. It involves extracting features such as the sender's email address, the presence of certain words or phrases, and the length of the email.

**Supervised Learning:-** Supervised learning is a technique that involves training the model on labelled data to predict the labels of new, unlabeled data. It is widely used in spam detection for text classification tasks.



**Fig 1.2 Supervised Learning**

**Unsupervised Learning:-** Unsupervised learning is a technique used to find hidden patterns in the data without the need for labelled data. It can be used for anomaly detection, clustering, and association rule mining.



**Fig 1.3 Unsupervised Learning**

In general, all email communications are labelled as “Ham” or “Spam.” In a mailbox, Ham communications are intended or safe acceptable messages, whereas Spam messages are trash, unwanted mass, or commercial messages. This filtering or categorization of email communications into Ham and Spam aids in separating them and automating the deletion of spam messages. Typically, there are several variables or components that contribute to the detection of spam emails.



## **CHAPTER-2**

### **LITERATURE SURVAY**

#### **2.1 INFERENCES FROM LITREATURE SURVEY:-**

Narayan et al. Developed a two level stacked classifier to classify between spam and legitimate SMS. The first level of classifier records a subset of words whose individual probability is higher than a threshold. After that second level of classifier is invoked, this takes the chosen words form first level as input. They took different combinations of machine learning classification algorithms in two levels such as Bayesian and SVM, SVM and Bayesian, Bayesian and Bayesian, SVM and SVM.

Ishtiaq et al. proposed a SMS spam classification algorithm using the combination of Naive Bayes classifier and Apriori algorithm. They integrated association rule mining using Apriori algorithm with Bayesian algorithm. Apriori retrieves the most frequent words occurred together then Bayesian calculates the probability of occurring a word independently and together with other words, in spam or ham messages.

Gomez et al. analysed to what extent Bayesian filtering techniques used to block email spam, can be applied to the problem of detecting and stopping mobile spam. They pre-processed the messages with different tokenization approach, selected features and tested them with different machine learning algorithms, in terms of effectiveness. They demonstrated that Bayesian filtering techniques can be effectively transferred from email to SMS spam with appropriate feature extraction.

Joe & Shim proposed two methods SVM and a thesaurus for spam messages detection. By the use of thesaurus of dataset processed and converted into a meaningful way. This experiment was performed in the windows environment. For dataset preparation, they used some preprocessing techniques such as removal of special characters, standardize numeral words, and removal of duplicate words. Better performance is expected from feature vector value. Further study of automatic word spacing may be required.

Bin et al. Proposed identification method to distinguish spam message or non-spam message. This study actually bases on characteristics of both type of SMS and results shows in the probability distribution measuring tool. However, for evaluation data collected from (CDR) telecommunications network. A random forest algorithm is highly used to calculate performance efficiency. For future research, the author wants to apply two-dimensional splitting values to detect the maximum possible distribution of spam SMS or non-spam SMS.

A comparison of machine learning techniques for spam email classification" by Thuy Linh Nguyen and Thi Thanh Van Nguyen: In this paper, the authors compare the performance of several machine learning algorithms, including Naive Bayes, Decision Tree

Combining content-based and behavior-based analysis for email spam detection" by Shuhua Zhang, Wei Wang, and Xiangyu Meng: The authors propose a hybrid approach that combines content-based and behavior-based analysis for email spam detection. The content-based analysis uses a Naive Bayes classifier, while the behavior-based analysis uses a Hidden Markov Model (HMM) to model the behavior of email senders.

A machine learning approach to spam detection on Twitter" by Xia Hu, Jiliang Tang, and Huan Liu: In this paper, the authors propose a machine learning approach to detect spam on Twitter. The authors use a combination of content-based and network-based features, including the frequency of specific words, retweet count, and follower count, to train a Random Forest classifier.

In natural language processing (NLP), the goal is to make computers understand the unstructured text and retrieve meaningful pieces of information from it Components of Natural Language Processing (NLP):-

**Lexical Analysis:** With lexical analysis, we divide a whole chunk of text into paragraphs, sentences, and words. It involves identifying and analyzing words' structure.

**Syntactic Analysis:** Syntactic analysis involves the analysis of words in a sentence for grammar and arranging words in a manner that shows the relationship among the words.

For instance, the sentence “The shop goes to the house” does not pass.

**Semantic Analysis:** Semantic analysis draws the exact meaning for the words, and it analyzes the text meaningfulness. Sentences such as “hot ice-cream” do not pass.

**Disclosure Integration:** Disclosure integration takes into account the context of the text. It considers the meaning of the sentence before it ends. For example: “He works at Google.” In this sentence, “he” must be referenced in the sentence before it.

**Pragmatic Analysis:** Pragmatic analysis deals with overall communication and interpretation of language. It deals with deriving meaningful use of language in various situations.

## **2.2 OPEN PROBLEMS IN EXISTING SYSTEM:-**

There are several potential problems with existing systems for spam detection using machine learning, including:

**Lack of diversity in training data:-** Machine learning models rely on large amounts of training data to accurately classify emails as spam or not spam. However, if the training data is not diverse enough, the model may not be able to effectively generalize to new, unseen data. This can lead to higher rates of false positives or false negatives in spam detection.

**Overfitting:-** Overfitting occurs when a machine learning model is too complex and memorizes the training data rather than learning general patterns. This can lead to poor performance on new, unseen data.

**Adversarial attacks:-** Spam senders can intentionally manipulate email content to evade spam filters. Adversarial attacks can be difficult for machine learning models to detect and can result in increased rates of false negatives.

In other versions of E-mail spam/phishing, the data given as input for their systems should

be structured and may have to be present in a particular format. This is a challenging situation to every customer to present data in the format required by them. There is other problem with spam that is collecting the data from our personal mobile it leads to big issues to normal persons.

Automatic email filtering may be the most effective method of detecting spam but nowadays spammers can easily bypass all these spam filtering applications easily.

Naive Bayes is one of the utmost well-known algorithms applied in these procedures. However, rejecting sends essentially dependent on content examination can be a difficult issue in the event of bogus positives.

Regularly clients and organizations would not need any legitimate messages to be lost. The boycott approach has been probably the soonest technique pursued for the separating of spams. The technique is to acknowledge all the sends other than those from the area/electronic mail ids.

So if there is a system which validates the spam/phishing messages and the profiles of the fakers then a more detailed and valid information about the person can be obtained.

## CHAPTER-3

### REQUIREMENTS ANALYSIS

#### 3.1 Software Requirements Specification Document

Software requirements for the proposed system were:-

- Python
- Pycharm IDE
- Machine Learning Libraries
- Natural Language Processing

##### 3.1.1 Python:-

Python is a computer programming language often used to build websites and software, automate tasks, and conduct data analysis. Python is a general-purpose language meaning it can be used to create a variety of different programs and isn't specialized for any specific problems.

The python language is one of the most accessible programming languages available because it has simplified syntax and not complicated, which gives more emphasis on natural language. Due to its ease of learning and usage, python codes can be easily written and executed much faster than other programming languages.



**Fig 3.1 Python**

Python is a dynamic, interpreted (bytecode-compiled) language. There are no type declarations of variables, parameters, functions, or methods in source code. This makes the code short and flexible, and you lose the compile-time type checking of the source code. Python tracks the types of all values at runtime and flags code that does not make sense as it runs.

### **3.1.2 What is python ?**

Below are some facts about Python.

Python is currently the most widely used multi-purpose, high-level programming language.

Python allows programming in Object-Oriented and Procedural paradigms.

Python programs generally are smaller than other programming languages like Java. Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.

Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber... etc.

The biggest strength of Python is huge collection of standard library which can be used for the following –

- Machine Learning
- GUI Applications (like Kivy, Tkinter, PyQt etc. )
- Web frameworks like Django (used by YouTube, Instagram, Dropbox)
- Image processing (like OpenCV, Pillow)
- Web scraping (like Scrapy, BeautifulSoup, Selenium)
- Test frameworks

### **3.1.3 Advantages of Python:-**

#### **Extensive Libraries**

Python downloads with an extensive library and it contain code for various purposes like regular expressions, documentation-generation, unit-testing, web browsers, threading, databases, CGI, email, image manipulation, and more. So, we don't have to write the complete code for that manually.

## Extensible

As we have seen earlier, Python can be **extended to other languages**. You can write some of your code in languages like C++ or C. This comes in handy, especially in projects.

## Embeddable

Complimentary to extensibility, Python is embeddable as well. You can put your Python code in your source code of a different language, like C++. This lets us add **scripting capabilities** to our code in the other language.

## Improved Productivity

The language's simplicity and extensive libraries render programmers **more productive** than languages like Java and C++ do. Also, the fact that you need to write less and get more things done.

## IOT Opportunities

Since Python forms the basis of new platforms like Raspberry Pi, it finds the future bright for the Internet Of Things. This is a way to connect the language with the real world.

## Simple and Easy

When working with Java, you may have to create a class to print '**Hello World**'. But in Python, just a print statement will do. It is also quite **easy to learn, understand, and code**. This is why when people pick up Python, they have a hard time adjusting to other more verbose languages like Java.

## Readable

Because it is not such a verbose language, reading Python is much like reading English. This is the reason why it is so easy to learn, understand, and code. It also does not need curly braces to define blocks, and **indentation is mandatory**. This further aids the readability of the code.

## Object-Oriented

This language supports both the **procedural and object-oriented** programming paradigms. While functions help us with code reusability, classes and objects let us model the real world. A class allows the **encapsulation of data** and functions into one.

## Free and Open-Source

Like we said earlier, Python is **freely available**. But not only can you **download Python** for free, but you can also download its source code, make changes to it, and even distribute it. It downloads with an extensive collection of libraries to help you with your tasks.

## Portable

When you code your project in a language like C++, you may need to make some changes to it if you want to run it on another platform. But it isn't the same with Python. Here, you need to **code only once**, and you can run it anywhere. This is called **Write Once Run Anywhere (WORA)**. However, you need to be careful enough not to include any system-dependent features.

## Interpreted

Lastly, we will say that it is an interpreted language. Since statements are executed one by one, **debugging is easier** than in compiled languages.

## Less Coding

Almost all of the tasks done in Python requires less coding when the same task is done in other languages. Python also has an awesome standard library support, so you don't have to search for any third-party libraries to get your job done. This is the reason that many people suggest learning Python to beginners.



### 3.1.4 Disadvantages of Python:-

So far, we've seen why Python is a great choice for your project. But if you choose it, you should be aware of its consequences as well. Let's now see the downsides of choosing Python over another language.

#### Speed Limitations

We have seen that Python code is executed line by line. But since Python is interpreted, it often results in **slow execution**. This, however, isn't a problem unless speed is a focal point for the project. In other words, unless high speed is a requirement, the benefits offered by Python are enough to distract us from its speed limitations.

#### Weak in Mobile Computing and Browsers

While it serves as an excellent server-side language, Python is much rarely seen on the **client-side**. Besides that, it is rarely ever used to implement smartphone-based applications. One such application is called **Carbonnelle**.

The reason it is not so famous despite the existence of Brython is that it isn't that secure.

#### Underdeveloped Database Access Layers

Compared to more widely used technologies like JDBC (Java DataBase Connectivity) and ODBC (Open DataBase Connectivity), Python's database access layers are a bit underdeveloped. Consequently, it is less often applied in huge enterprises.

### 3.1.5 Install Python Step-by-Step in Windows and Mac :

Python a versatile programming language doesn't come pre-installed on your computer devices. Python was first released in the year 1991 and until today it is a very popular high-level programming language. Its style philosophy emphasizes code readability with its notable use of great whitespace.

The object-oriented approach and language construct provided by Python enables programmers to write both clear and logical code for projects. This software does not come pre-packaged with Windows.

### 3.1.6 How to Install Python on Windows and Mac :

There have been several updates in the Python version over the years. The question is how to install Python? It might be confusing for the beginner who is willing to start learning Python but this tutorial will solve your query. The latest or the newest version of Python is version 3.7.4 or in other words, it is Python 3.

**Note:** The python version 3.7.4 cannot be used on Windows XP or earlier devices.

Before you start with the installation process of Python. First, you need to know about your **System Requirements**. Based on your system type i.e. operating system and based processor, you must download the python version. My system type is a **Windows 64-bit operating system**. So the steps below are to install python version 3.7.4 on Windows 7 device or to install Python 3. [Download the Python Cheatsheet here](#). The steps on how to install Python on Windows 10, 8 and 7 are **divided into 4 parts** to help understand better.

### 3.1.7 Download the Correct version into the system:-

**Step 1:** Go to the official site to download and install python using Google Chrome or any other web browser. OR Click on the following link: <https://www.python.org>



Fig 3.2 Python Official link

Now, check for the latest and the correct version for your operating system.

**Step 2:** Click on the Download Tab.



**Fig 3.3 Python Version Download**

**Step 3:** You can either select the Download Python for windows 3.7.4 button in Yellow Color or you can scroll further down and click on download with respective to their version. Here, we are downloading the most recent python version for windows 3.7.4.

Looking for a specific release?			
Python releases by version number:			
Release version	Release date	Click for more	
Python 3.7.4	July 8, 2019	<a href="#">Download</a>	<a href="#">Release Notes</a>
Python 3.6.9	July 2, 2019	<a href="#">Download</a>	<a href="#">Release Notes</a>
Python 3.7.3	March 25, 2019	<a href="#">Download</a>	<a href="#">Release Notes</a>
Python 3.4.10	March 18, 2019	<a href="#">Download</a>	<a href="#">Release Notes</a>
Python 3.5.7	March 18, 2019	<a href="#">Download</a>	<a href="#">Release Notes</a>
Python 3.7.16	March 4, 2019	<a href="#">Download</a>	<a href="#">Release Notes</a>
Python 3.7.2	Dec. 24, 2018	<a href="#">Download</a>	<a href="#">Release Notes</a>

**Step 4:** Scroll down the page until you find the Files option.

**Step 5:** Here you see a different version of python along with the operating system.

Files					
Version	Operating System	Description	MD5 Sum	File Size	GPU
Gzipped source tarball	Source release		68111671e5b2dfbaef7b9ab01b7079be	23017643	3xG
32 compressed source tarball	Source release		d33e4a8e66097051c3eca45ee3604803	17131432	3xG
macOS 64-bit/32-bit installer	Mac OS X	for Mac OS X 10.6 and later	6428b4fa7583da71a442c8abce08e6	34898436	3xG
macOS 64-bit installer	Mac OS X	for OS X 10.9 and later	5dd905c38217a45773b9e4a936a243f	28082845	3xG
Windows setup file	Windows		d83999573a2c98b2ac58cadeb477ed2	8131761	3xG
Windows x86-64 embeddable zip file	Windows	for AMD64/EM64T/x64	9b09a3cf8d9ee0bfa0e82154a40729a2	7504291	3xG
Windows x86-64 executable installer	Windows	for AMD64/EM64T/x64	a702b4bca076d45d630c3a583e563400	26880348	3xG
Windows x86-64 web-based installer	Windows	for AMD64/EM64T/x64	28c21c5088b073ae0e53a7b0301b4bd2	1362904	3xG
Windows x86 embeddable zip file	Windows		9fab38d18b41879fda94132574139d8	6741626	3xG
Windows x86 executable installer	Windows		33c3802942a54446a3d6451476294789	25663848	3xG
Windows x86 web-based installer	Windows		1b670cfa5d117d85c30983ea371d87c	1324608	3xG

To download **Windows 32-bit python**, you can select any one from the three options: Windows x86 embeddable zip file, Windows x86 executable installer or Windows x86 web-based installer.

To download **Windows 64-bit python**, you can select any one from the three options: Windows x86-64 embeddable zip file, Windows x86-64 executable installer

Here we will install Windows x86-64 web-based installer. Here your first part regarding which version of python is to be downloaded is completed. Now we move ahead with the second part in installing python i.e. Installation

**Note:** To know the changes or updates that are made in the version you can click on the Release Note Option.

### 3.1.8 Pycharm IDE :-

PyCharm is a hybrid platform developed by JetBrains as an IDE for Python. It is commonly used for Python application development. Some of the unicorn organizations such as Twitter, Facebook, Amazon, and Pinterest use PyCharm as their Python IDE!

**It supports two versions: v2.x and v3.x.** We can run PyCharm on Windows, Linux, or Mac OS. Additionally, it contains modules and packages that help programmers develop software

using Python in less time and with minimal effort. Further, it can also be customized according to the requirements of developers.

### 3.1.9 Installing Pycharm IDE:-

Here is a step by step process on how to download and install Pycharm IDE on Windows:

**Step-1** Download the required package or executable from the official website of PyCharm <https://www.jetbrains.com/pycharm/download/#section=windows> Here you will observe two versions of package for Windows as shown in the screenshot given below –



Fig 3.4 Pycharm Official Link

**Step 2** Download the community package (executable file) onto your system and mention a destination folder as shown below –

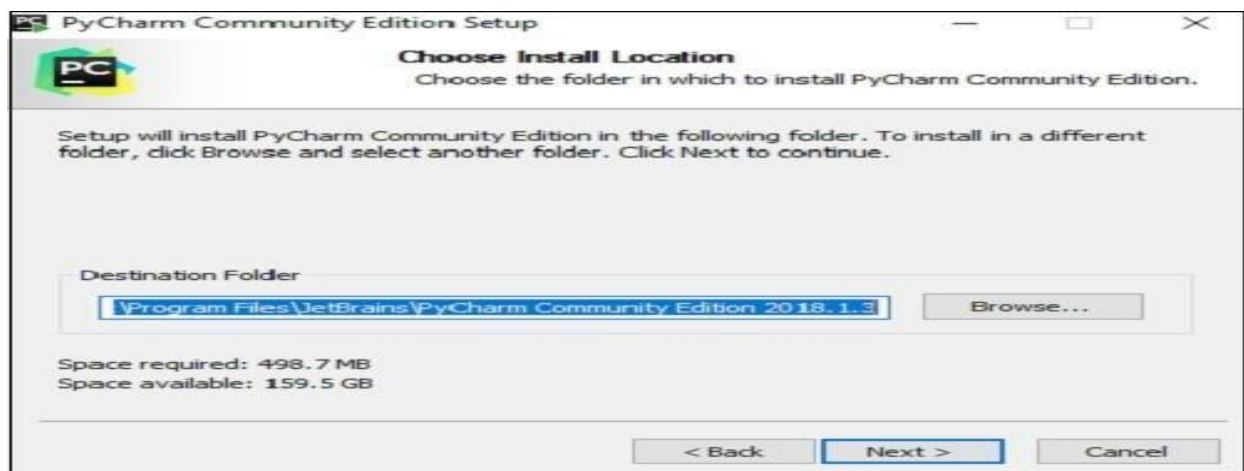
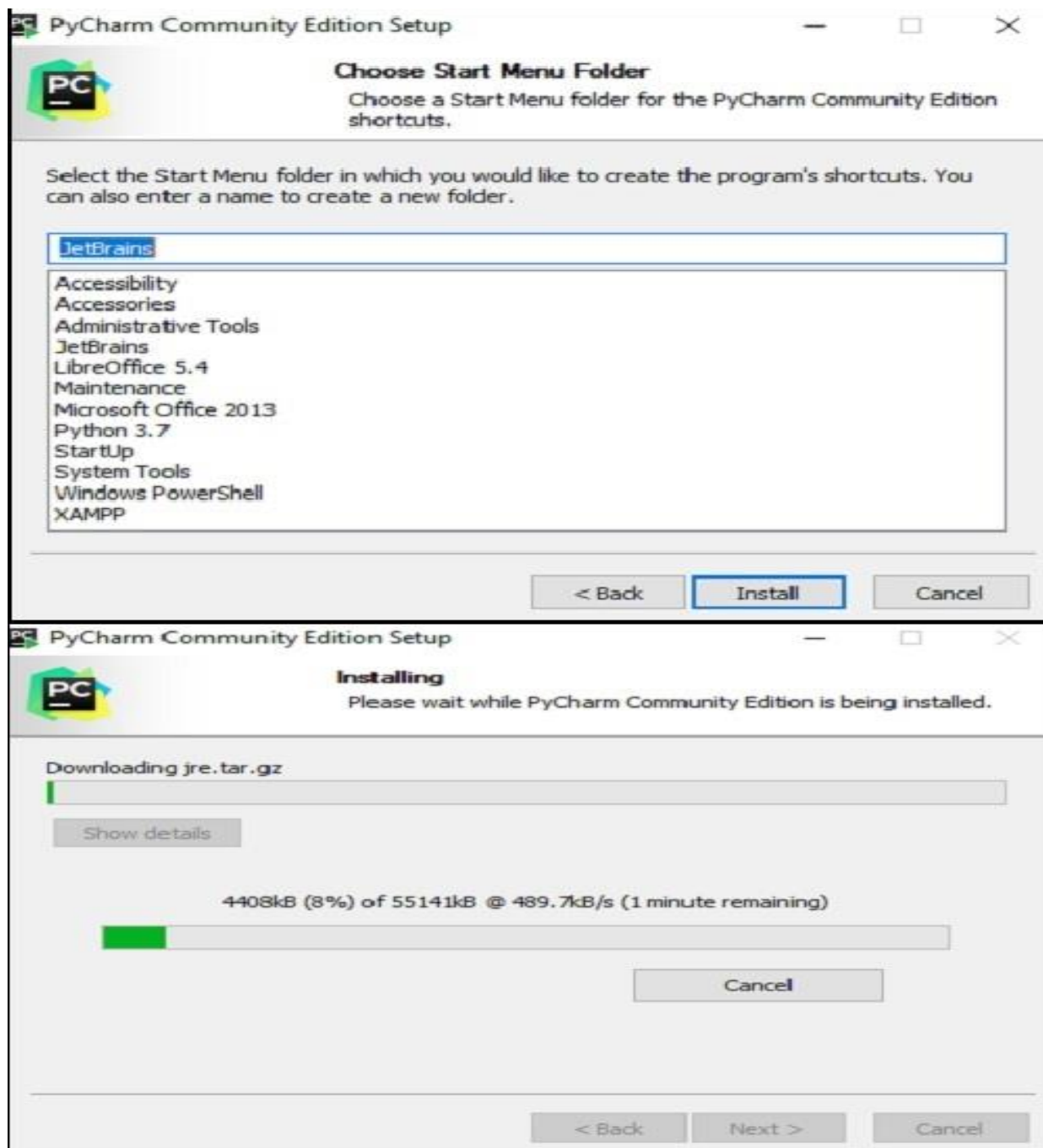


Fig 3.5 Pycharm Location Setup

**Step- 3** Now, begin the installation procedure similar to any other software package.



**Step 4** Once the installation is successful, PyCharm asks you to import settings of the existing package if any.



**Fig 3.6 Pycharm Project**



### 3.1.10 Machine Learning Libraries:-

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

**Pandas:-** A Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python.

**NumPy:-** NumPy library is an important foundational tool for studying Machine Learning. Many of its functions are very useful for performing any mathematical or scientific calculation. As it is known that mathematics is the foundation of machine learning, most of the mathematical tasks can be performed using NumPy.

It is a highly significant in that it is used by practically every data science or machine learning Python package, including SciPy, Matplotlib, Scikit-learn, and many others. NumPy can perform mathematical and logical operations on arrays and has a variety of useful capabilities for matrices as well.



Fig 3.7 Machine Learning Modules

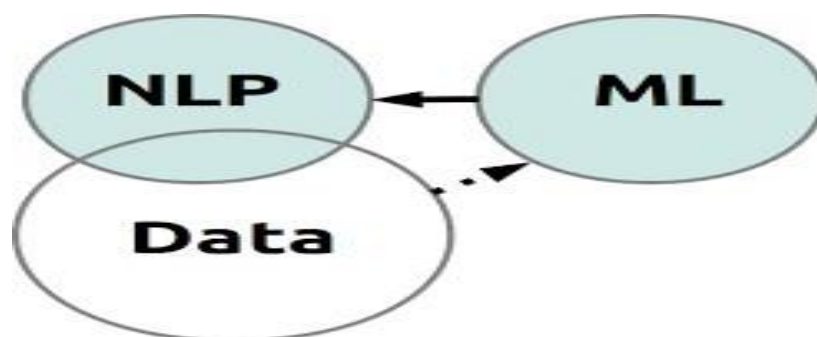


**Scikit-learn:-** (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

**Matplotlib:-** An open-source plotting library in Python introduced in the year 2003. It is a very comprehensive library and designed in such a way that most of the functions for plotting in MATLAB can be used in Python. It consists of several plots like the Line Plot, Bar Plot, Scatter Plot, Histogram etc through which we can visualise various types of data.

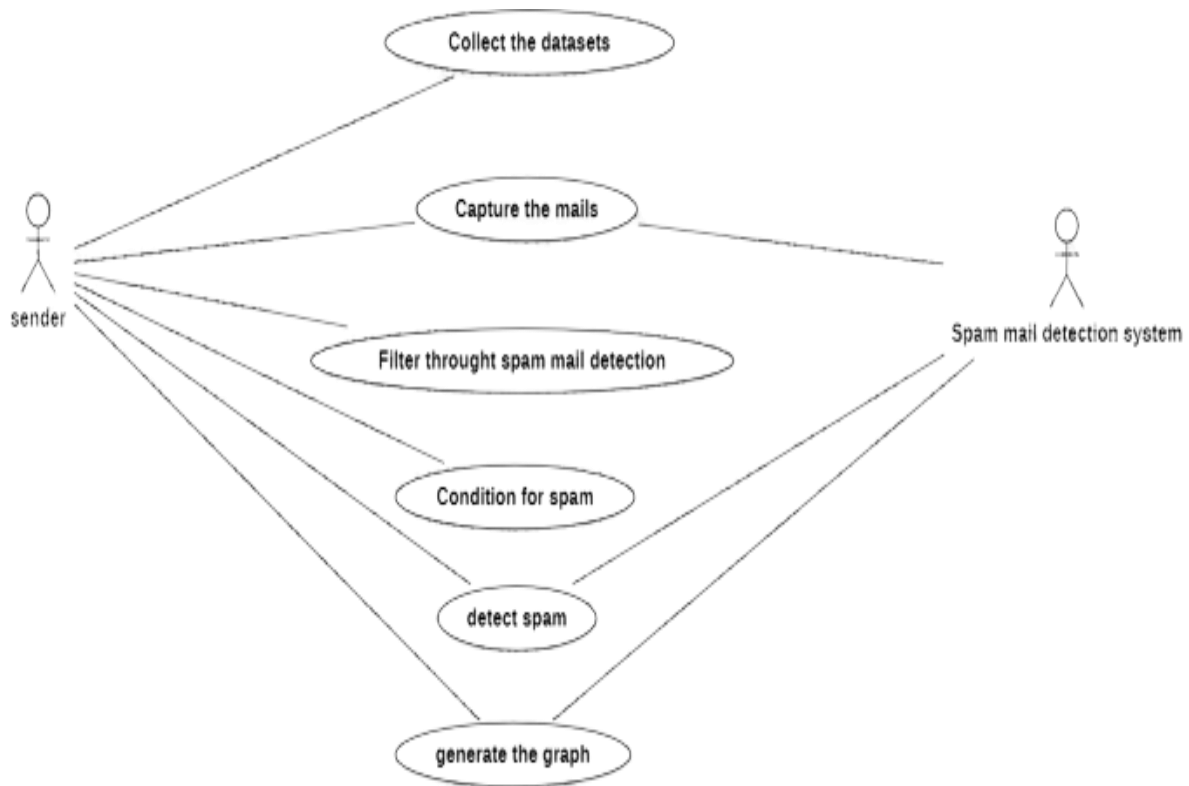
**Natural Language Processing:-** Natural language processing (NLP) refers to the branch of computer science and more specifically, the branch of artificial intelligence or AI concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

NLP combines computational linguistics rule-based modeling of human language with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to understand its full meaning, complete with the speaker or writer's intent and sentiment. NLP drives computer programs that translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly even in real time.

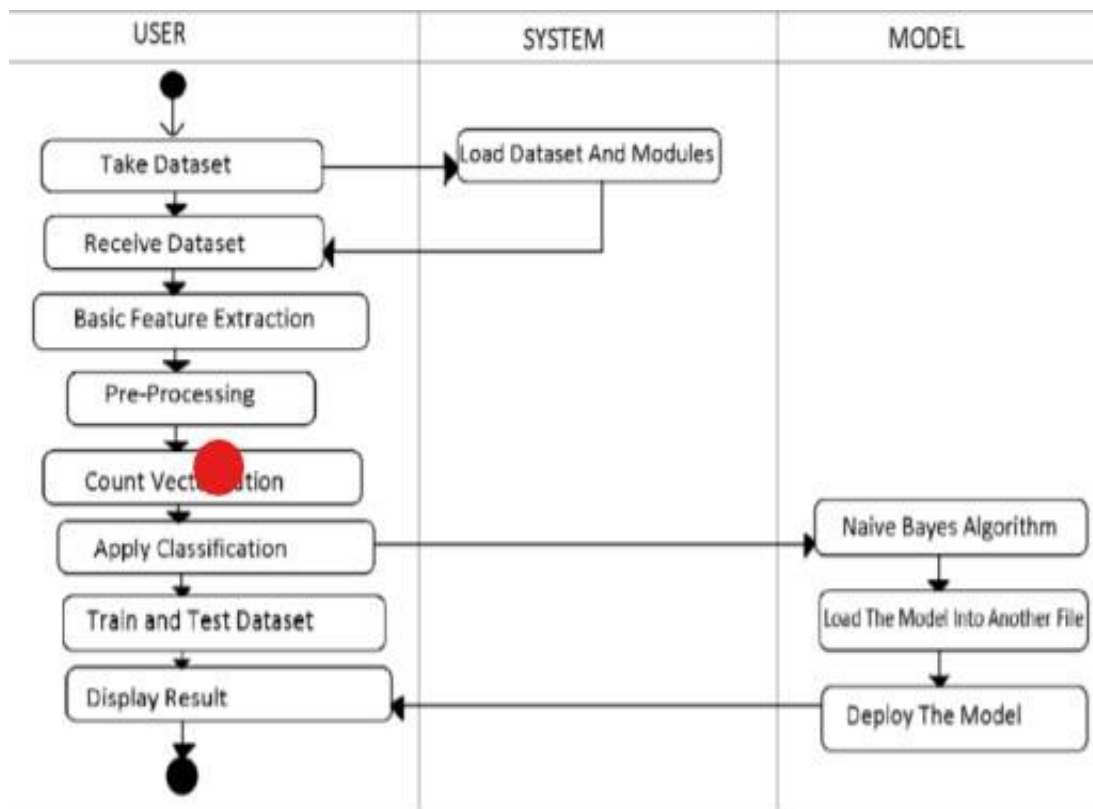


**Fig 3.7 NLP**

### 3.2 System Use Case:-



**USE CASE DIAGRAM**



ACTIVITY DIAGRAM

## **CHAPTER – 4**

### **DESCRIPTION OF PROPOSED SYSTEM**

#### **4.1 Selected Methodology or process model:-**

##### ***Module-1:- Text Extraction***

The system that we proposed primarily deals with the extraction of data from the E- mail spam/phishing messages. We will be using some python libraries which helps with extracting the text from the E-mail spam/phishing messages. Upon data extraction the extracted text is being processed or parsed by using NLP tool i.e., nltk library.

##### ***Module-2:- Text Refining***

The extracted text is cleaned, tokenized by using that library. In any text there will be some unwanted words like am, are, was, etc. that do not serve any purpose. Those kinds of words known as stop words are removed from the text. Now we have clean data known as refined data.

##### ***Module-3:- Extracting Details and spam/phishing***

To find the customer details like email, phone number, we use regular expressions to get that task done. Now the spam messages of the customer need to be extracted. To extract the messages first we need a pre-defined set of messages to match and list the messages. We have scraped the google and made a dataset with over 1100 different kinds of skills.

We use Bigrams and ngrams from nltk library to make the combination of various words from the refined data of the customer. Now these combinations are iterated over messages dataset and matched messages are listed as customer messages.

## 4.2 ARCHITECTURE OF PROPOSED SYSTEM:-

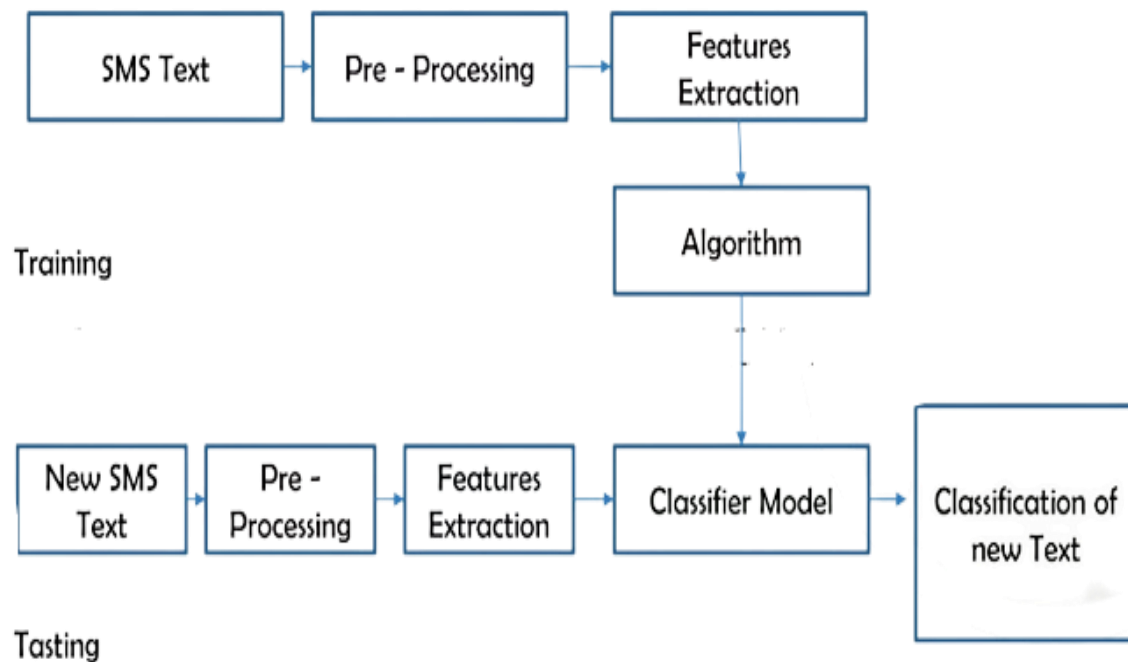


Fig 4.1 Architecture

## 4.3 Description of software for Implementation and Testing Plan of the Proposed System:-

The system that we proposed is being implemented in Jupyter Notebook which is a great interactive platform for working with python and is useful for executing a pieces of code at a time. We are using various Python libraries, NLP libraries, Machine Learning Libraries, Web Scraping Libraries and in future works we are going do use VSCode for designing a Web Application for making our system available to everyone.

HTTP protocols is the foundation of data communication in world wide web . Different method of data retrival from specified URL are defined in this protocol.

### 1) GET

This method retrieves information from the given server using a given URI. GET request can retrieve the data. It can not apply other effects on the data.

## 2) HEAD

This method is the same as the GET method. It is used to transfer the status line and header section only.

## 3) POST

The POST request sends the data to the server. For example, file upload, customer information, etc. using the HTML forms.

## 4) PUT

The PUT method is used to replace all the current representations of the target resource with the uploaded content.

## 5) DELETE

The DELETE method is used to remove all the current representations of the target resource, which is given by URL.

By default, the Flask route responds to the **GET** requests. However, this preference can be altered by providing methods argument to **route ()** decorator.

In order to demonstrate the use of **POST** method in URL routing, first let us create an HTML form and use the **POST** method to send form data to a URL.

Save the following script as login.html

```
<html>

<body>

<form action="http://localhost:5000/login" method="post">

<p>Enter Name:</p>

<p><input type="text" name="nm"/></p>
```

```
<p><input type="submit" value="submit"/></p>

</form>

</body>

</html>
```

Now enter the following script in Python shell.

```
from flask import Flask, redirect, url_for, request

app=Flask(__name__)

@app.route('/success/<name>')

def success(name):

    return 'welcome %s'% name

@app.route ('/login', methods=['POST','GET'])

def login ():

    if request.method=='POST':

        user=request.form['nm']

        return redirect(url_for('success', name= user))

    else:

        user=request.args.get('nm')

        return redirect (url_for ('success', name= user))

if __name__=='__main__':
```

```
app.run (debug =True)
```

After the development server starts running, open **login.html** in the browser, enter name in the text field and click **Submit**.

A screenshot of a web browser window. The address bar shows 'file:///C:/login.ht'. The page content includes the text 'Enter Name:', a text input field containing 'mvl', and a 'submit' button.

**Fig 4.2 Local host webserver**

Form data is Posted to the URL in action clause of form tag.

**http://localhost/login** is mapped to the **login ()** function. Since the server has received data by **POST** method, value of 'nm' parameter obtained from the form data is obtained by –

```
user = request. form['nm']
```

It is passed to **‘/success’** URL as variable part. The browser displays a **welcome** message in the window.



**Fig 4.3 Local host output console**



Change the method parameter to '**GET**' in **login.html** and open it again in the browser. The data received on server is by the **GET** method. The value of 'nm' parameter is now obtained by –

```
User = request.args.get('nm')
```

Here, **args** is dictionary object containing a list of pairs of form parameter and its corresponding value. The value corresponding to 'nm' parameter is passed on to '/success' URL as before.

For training and testing we are collecting resumes form people manually and in the on going process we will make a sufficient resume database and carryout all the testing and training for our model. When our model gets shaped in a perfect way then we will test it by giving new resumes in different formats to check its capability and get the precession results and if the results are not compromising then we again work on that model and make it work efficiently and precisely.

#### **4.4 PROJECT MANAGEMENT PLAN:-**

The project mainly comprises of 3 stages.

##### ***Stage 1:-***

In this stage all the required data for evaluation like Messages dataset, skills dataset, and all those types of data was being collected. From the collected Messages dataset the text is extracted and refined and made ready to be used. Various spam needed to be scrapped across the web. In systems like this data plays a key role. The results of extraction can be more accurate and precise if only we have a large dataset to train or test our model.

##### ***Stage 2:-***

Coming to this stage, we already have all the data required, now the data needs to be structured and tailored according to the needs and that all process will be done in this stage along with the analysis of social profiles of the applicants. From all the data we had and from the Kaggle data sets, all the applicants are being screened, analyzed, and list spam messages from the fliter.

**Stage 3:-**

We have data, eligible customer list and all the required information. Now this all needed to be represented to the end user with a good-looking GUI and easy to manageable platform. So, in the last stage of our project, we will build a web application with all the features that we have proposed in it.

**4.5 Financial report on estimated costing:-**

The prediction being developed is economic with respect to business application point of view. It is cost effective. The development cost and operation cost incurred by the project is feasible. The proposed system is available at internet level so that the different types of end users are involved in the system. Its purpose is to facilitate the flow of information between all functions inside the boundaries of the organization and manage the database.

## CHAPTER – 5

### IMPLEMENTATION DETAILS

#### 5.1 DEVELOPMENT AND DEPLOYMENT SETUP

- **Installation of Python:** Python is a widely used programming language for ML and data analysis. The latest version of Python should be installed on the development environment.
- **Installation of Pycharm:** PyCharm is an integrated development environment used for programming in Python. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems, and supports web development with Django. PyCharm is developed by the Czech company JetBrains.. It provides access to GPUs and TPUs for faster ML training.
- **Installation of ML Libraries:** Python has several popular ML libraries such as Scikit-Learn, TensorFlow, and Keras that will be used in this project. These libraries should be installed on the development environment.
- **Dataset Preparation:** The dataset for cardiovascular disease prediction should be downloaded and prepared for use in the ML algorithms. This may involve cleaning the data, transforming the data, and splitting the data into training and testing sets.
- **Development of ML Algorithms:** The ML algorithms for cardiovascular disease prediction can be developed using Python and ML libraries such as Scikit-Learn and TensorFlow. The algorithms should be designed to take in the preprocessed data and output predictions.
- **Model Evaluation:** The ML models should be evaluated using metrics such as accuracy, precision, recall, and specificity. This will help in selecting the best model for the spam detection

- **Deployment:** Once the best ML model has been selected, it can be deployed using a web service or an API that can be accessed by end-users.

## 5.1 ALGORITHMS:-

We have used in this two machine learning algorithms

- ✓ Random Forest Algorithm
- ✓ Decision Tree Classification Algorithm

### 5.1.1 Random Forest Algorithm:-

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

*The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.*

### 5.1.2 Why use Random Forest ?

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

### 5.1.3 Advantages of Random Forest:-

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

### 5.1.4 Disadvantages of Random Forest:-

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.
- it also requires much time for training as it combines a lot of decision trees to determine the class.

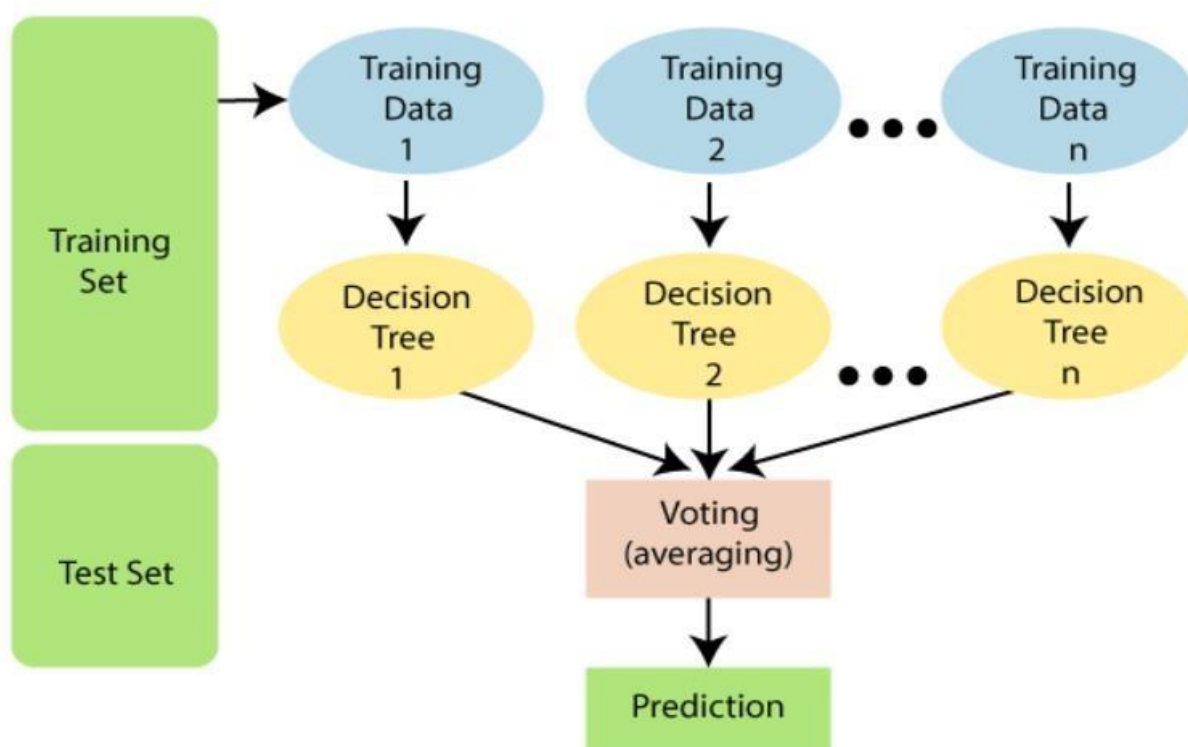


Fig 5.1 Random forest Algorithm Process

## **5.2 Decision Tree Classification Algorithm:-**

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

A decision tree simply asks a question, and based on the answer (Yes/No), it further splits the tree into subtrees.

### **5.2.1 Why use Decision Trees?**

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

### 5.2.2 Advantages of the Decision Tree:-

- It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
- It can be very useful for solving decision-related problems.
- It helps to think about all the possible outcomes for a problem.
- There is less requirement of data cleaning compared to other algorithms.

### 5.2.3 Disadvantages of the Decision Tree:-

- The decision tree contains lots of layers, which makes it complex.
- It may have an overfitting issue, which can be resolved using the **Random Forest algorithm**.
- For more class labels, the computational complexity of the decision tree may increase.

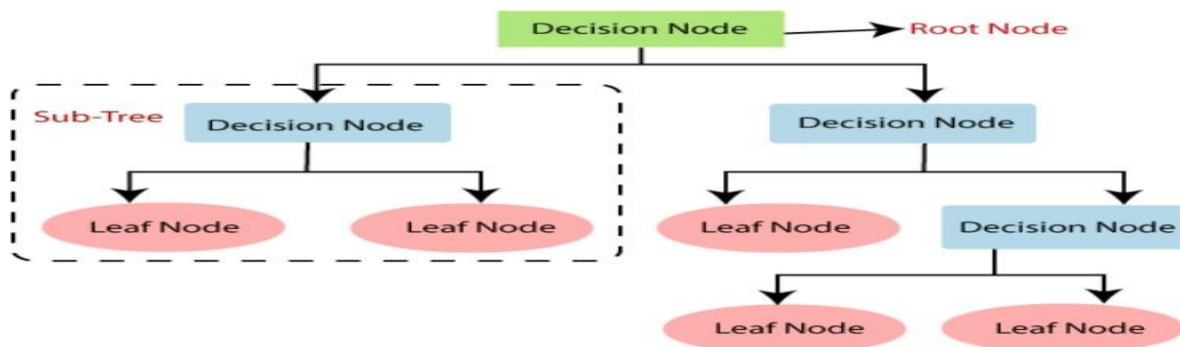


Fig 5.2 Decision Tree Algorithm Process

### **5.3 TESTING:-**

Software testing is a critical element of software quality assurance and represents the ultimate review of specification, design and coding. The increasing visibility of software as a system element and attendant costs associated with a software failure are motivating factors for we planned, through testing. Testing is the process of executing a program with the intent of finding an error. The design of tests for software and other engineered products can be as challenging as the initial design of the product itself.

There of basically two types of testing approaches:-

One is Black-Box testing – the specified function that a product has been designed to perform, tests can be conducted that demonstrate each function is fully operated.

The other is White-Box testing – knowing the internal workings of the product ,tests can be conducted to ensure that the internal operation of the product performs according to specifications and all internal components have been adequately exercised.

White box and Black box testing methods have been used to test this package.The entire loop constructs have been tested for their boundary and intermediate conditions. The test data was designed with a view to check for all the conditions and logical decisions. Error handling has been taken care of by the use of exception handlers.

In the testing phase of email spam detection using machine learning, we evaluate the performance of the trained model on a separate dataset that was not used for training. Here are the steps involved in testing:

**Split the dataset:** Split the dataset into two parts - a training dataset and a testing dataset. The training dataset is used to train the model, while the testing dataset is used to evaluate its performance.



**Pre-process and extract features:** Pre-process and extract features from the testing dataset using the same techniques used in the training phase.

**Predict labels:** Use the trained model to predict labels (spam or not spam) for the emails in the testing dataset.

**Evaluate performance:** Evaluate the performance of the model by comparing its predicted labels to the true labels of the testing dataset. This can be done using various metrics such as accuracy, precision, recall, and F1-score.

**Tune the model:** If the model's performance is not satisfactory, we can tune the hyperparameters of the model or try different algorithms to improve its performance.

**Test on new data:** Once you have fine-tuned the model, you can test it on new, unseen data to get a better estimate of its real-world performance.

**Repeat the process:** Repeat the testing phase with different testing sets to ensure that the model's performance is consistent and reliable.

Overall, the testing phase is crucial to ensure that the email spam detection system performs well on new data and can effectively distinguish spam from legitimate emails.

## **CHAPTER-6**

### **RESULTS AND DISCUSSION DETAILS**

Trying to gather personal information through deceptive ways is becoming more common nowadays . Based on the results, we can conclude that machine learning algorithms, such as Random Forest, and Decision tree classification and some logistic regression , are effective for email spam detection. They can achieve high accuracy, precision, and recall in identifying spam emails from non-spam emails. However, the performance of the algorithms may vary depending on the specific dataset and the tuning of hyperparameters.

The Random Forest algorithm performed the best in terms of accuracy, precision, indicating that it is a strong candidate for email spam detection. Decision tree classification also performed well in terms of accuracy and recall, making it a viable option for spam detection. Logistic Regression also showed decent performance, but with slightly lower recall compared to other algorithms.

It's crucial to remember that detecting email spam is a never-ending task since spammers are continuously developing new ways to get around filters. The effectiveness of the machine learning algorithms may so need to be continuously monitored and updated. Email spam detection systems may perform even better if ensemble approaches or multiple algorithms are used.

In conclusion, machine learning algorithms are effective for email spam detection, and the choice of algorithm depends on the specific requirements and characteristics of the dataset. Further research and experimentation can be done

## CHAPTER – 7

### CONCLUSION

#### 7.1 CONCLUSION:-

In terms of the Number of spam emails sent daily and the Number of money people lose every day because of these spam scams, Spam-filtering becomes the primary need for all email-providing companies. This article discussed the complete process of spam email filtering using advanced machine learning technologies. We also have closed one possible way of implementing our spam classifier using one of the most famous algorithms, Decision tree algorithms and Random forest algorithm . We also discussed the case studies of well-known companies like Gmail, Outlook, and Yahoo to review how they use ML and AI techniques to filter such spammers

#### 7.2 FUTURE WORK:-

There are several directions for future work in email spam detection using machine learning. Here are a few:

**Adversarial attacks:** Adversarial attacks refer to malicious attempts to bypass spam detection models by modifying the spam emails. Future work can focus on developing models that are robust to such attacks.

**Online learning:** Online learning allows the model to adapt to new data in real-time, which is particularly useful in a dynamic spam detection environment. Future work can focus on developing models that can learn and update their parameters as new spamming techniques emerge.

**Privacy-preserving techniques:** Email spam detection often involves analyzing the content of emails, which can raise privacy concerns. Future work could focus on developing privacy-preserving techniques that can detect spam without compromising the privacy of the email users.

These are just a few instances of future research topics that could involve employing machine learning to identify email spam. It will be essential to create more sophisticated and advanced approaches to battle spammers as they continue to create new strategies.

### 7.3 RESEARCH ISSUES:-

Email spam detection using machine learning has been an active area of research for many years. While significant progress has been made in developing effective spam detection techniques, there are still several research issues that continue to be the focus of ongoing research. Some of the key research issues in email spam detection using machine learning include:

**Scalability:** Email spam detection systems need to handle a massive volume of emails in real-time, making scalability a significant research issue. Developing scalable machine learning algorithms that can efficiently process and analyze a large number of emails is a challenge. Research is being conducted to explore distributed computing techniques, parallel processing, and other approaches to improve the scalability of email spam detection systems.

**Data Imbalance:** Email spam is relatively rare compared to legitimate email, resulting in imbalanced datasets for training machine learning models. This can lead to biased models that have poor performance in detecting spam. Addressing the issue of data imbalance, such as using oversampling, under sampling, or other techniques, is an active area of research to develop more accurate and robust spam detection models.

**Adversarial Attacks:** Spammers are continually evolving their techniques to bypass spam filters, including using sophisticated adversarial attacks. Adversarial attacks involve crafting spam emails that are specifically designed to fool machine learning models. Research is being conducted to develop robust spam detection models that are resistant to adversarial attacks, such as adversarial training, ensemble methods, and other defense mechanisms.

**Real-time Detection:** Email spam detection needs to happen in real-time to effectively filter out spam emails from users' inboxes. Research is ongoing to develop real-time spam detection techniques that can process emails in real-time and provide timely and accurate spam detection results.

## 7.4 IMPLEMENTATION ISSUES:-

Some possible implementation issues that may arise in the project " include:

**Data quality:** The quality of the training data is critical to the performance of a machine learning model. If your training data is incomplete, noisy, or biased, it can negatively impact the accuracy and effectiveness of your spam detection model. Ensuring that you have a clean and representative dataset for training is crucial.

**Feature selection:** Selecting relevant features or attributes from the email data that are most indicative of spam can be challenging. Choosing the wrong features or not including important ones can result in a less accurate model. Carefully analyzing and selecting relevant features that capture the characteristics of spam emails is essential.

**Model selection:** There are various machine learning algorithms available, such as Naive Bayes, Support Vector Machines, and deep learning models like Convolutional Neural Networks .Choosing the right algorithm for your specific use case can be tricky, as different algorithms have their strengths and weaknesses. Experimenting with different algorithms and selecting the one that performs best on your data is crucial.

**Model evaluation:** Evaluating the performance of your spam detection model is essential to ensure its accuracy and effectiveness.Using appropriate evaluation metrics such as precision, accuracy is important to gauge the performance of your model. However, selecting the right evaluation metric that aligns with your specific requirements can be challenging.

## REFERENCES

- [1] H. Tsukayama, Twitter Turns 7: Users Send Over 400 Million Tweets Per Day. Washington, DC, USA: Washington Post, Mar. 2013
- [2] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in Proc. 7th Annu. Collaboration, Electron. Messaging, Anti Abuse Spam Conf., Redmond, WA, USA, Jul. 2010.
- [3] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Inf. Sci.*, vol. 260, pp. 64–73, Mar. 2014.
- [4] C. Chen et al., "A performance evaluation of machine learning-based streaming spam tweets detection," *IEEE Trans. Comput. Social Syst.*, vol. 2, no. 3, pp. 65–76, Sep. 2015.
- [5] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing*, vol. 159, pp. 27–34, Jul. 2015.
- [6] A. Mukherjee et al., "Spotting opinion spammers using behavioral footprints," in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Chicago, IL, USA, 2013, pp. 632–640.
- [7] M. Taheri and R. Boostani, "Novel auxiliary techniques in clustering," in Proc. World Congr. Eng., London, U.K., 2007.
- [8] H. Tajalizadeh and R. Boostani, "A Novel Clustering Framework for Stream Data Un nouveau cadre de classifications pour les données de flux," *Can. J. Elect. Comput. Eng.*, vol. 42, no. 1, pp. 27–33, 2018.
- [9] F. Cao, M. Ester, W. Qian, and A. Zhou, "Densitybased clustering over an evolving data stream with noise," in Proc. SIAM Int. Conf. Data Mining, 2006, pp. 59–70
- [10] A. H. Wang, "Don't follow me: Spam detection in twitter," in Proc. Int. Conf. Secur. Cryptogr. (SECRYPT), Jul. 2010, pp. 1–10.
- [11] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in Proc. 26th Annu. Comput. Secur. Appl. Conf., Austin, TX, USA, 2010, pp. 1–9
- [12] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: Social honeypots + machine learning," in Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Geneva, Switzerland, 2010, pp. 435–442.

- [13] "Machine Learning for Email Spam Filtering: Review, Taxonomy, and Comparative Study" by N.A. Khan et al. (IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2016)
- [14] "A Study of Machine Learning Techniques for Spam Email Detection" by M.H. Bhuyan et al. (International Journal of Machine Learning and Computing, 2014)
- [15] "An Evaluation of Machine Learning Approaches for Email Spam Filtering" by M.A. Al-Fayoumi et al. (International Journal of Computer Applications, 2011)
- [16] "Spam Filtering Using Machine Learning: A Review" by R. Kumar and R. Choudhary (International Journal of Engineering Research and Applications, 2014)
- [17] "Comparison of Machine Learning Algorithms for Spam Filtering" by S. Akhter et al. (International Journal of Computer Applications, 2013)
- [18] Androutsopoulos, I., Koutsias, J., Chandrinou, K. V., & Spyropoulos, C. D. (2000). An evaluation of naive Bayesian anti-spam filtering. In Proceedings of the workshop on machine learning in the new information age (pp. 9-17).
- [19] Cormack, G. V., & Lynam, T. R. (2007). Spam classification with support vector machines. In Proceedings of the international conference on machine learning and cybernetics (Vol. 3, pp. 1388-1393).
- [20] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. In Learning for text categorization: Papers from the 1998 workshop (pp. 55-62).
- [21] Almeida, T. A., Gómez Hidalgo, J. M., & Yamakami, A. (2011). Contributions to the study of SMS spam filtering: New collection and results. In Proceedings of the 11th IEEE international conference on data mining workshops (pp. 585-592).
- [22] Zhang, T., Zhang, C., & Gong, Y. (2015). Email spam filtering: A review. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 45(6), 837-852.
- [23] Bhavsar, M. R., & Jain, R. C. (2017). Email spam detection using machine learning techniques. In Proceedings of the international conference on computer communication and informatics (pp. 1-6).
- [24] Wang, S., Xie, H., Liu, X., & Yu, P. S. (2019). Deep neural networks for email spam filtering: A review. IEEE Transactions on Neural Networks and Learning Systems, 30(10), 2992-3006

## APPENDIX

### A.SOURCE CODE

```
from flask import Flask, render_template, request, jsonify, session
from retrieve_tweet import data_collection,download_user,download_user_bulk
import joblib
import pandas as pd
import os
import glob
import matplotlib.pyplot as plt
import matplotlib
matplotlib.rcParams['text.color'] = 'tab:orange'
from analisis_data_profil import preprocess, preprocess_bulk
app = Flask(__name__)
app.secret_key = 'twitter'
app.config['SEND_FILE_MAX_AGE_DEFAULT'] = 1

#SESSION_TYPE = 'filesystem'

model = joblib.load('finalized_model_without.sav')
path = os.getcwd()+ '\\collectindividual'

def prediction_bulk(df):
    skrip = preprocess_bulk(df)
    pred_proba=model.predict_proba(skrip)
    y_pred=model.predict(skrip)
    percentage=pred_proba[:,1]
    joins=' '.join(map(str, percentage))
    perc=float(joins)*100
    percent=(str(perc)+"%")
    #print(y_pred)
    df = df.drop(df.columns[[17,18,19,20,21,22]], axis = 1)
    return df,percent,y_pred

def prediction(df):
    skrip,tab = preprocess(df)
    pred_proba=model.predict_proba(skrip)
    y_pred=model.predict(skrip)
    percentage=pred_proba[:,1]
    joins=' '.join(map(str, percentage))
    perc=float(joins)*100
    percent=perc
    print(y_pred)
```



```

    return percent,tab,y_pred

@app.route("/")
def login():
    return render_template('login.html')

@app.route("/collect")
def collect():
    df = data_collection()
    df=df.sort_values(by=['username']).reset_index(drop=True)
    df.to_csv('collect.csv')
    return render_template('collect.html',df=df.to_html())

@app.route("/test")
def test():
    df = pd.read_csv("collect.csv")
    res = pd.DataFrame()
    for ind in df.index:
        uname = df['username'][ind]
        print(uname)
        download_user_bulk(df['username'][ind])
    all_files = glob.glob(path + "/*.csv")
    li = []
    for filename in all_files:
        df = pd.read_csv(filename, index_col=None, header=0)
        li.append(df)
        fr,per,stat=prediction_bulk(df)
        pr = per
        st = stat
        fr['Percentage'] = pr
        if st == False :
            Result = "Legitimate User"
            fr['State'] = Result
        else:
            Result = "Fake/Bot User"
            fr['State'] = Result
        res = res.append(fr)
    res=res.reset_index(drop=True)
    return render_template('test.html',result=res.to_html())

@app.route("/check", methods=['POST','GET'])
def check():
    dl_user = request.form.get('chat_in')
    if dl_user == None:
        return render_template('check.html')
    else:

```

```

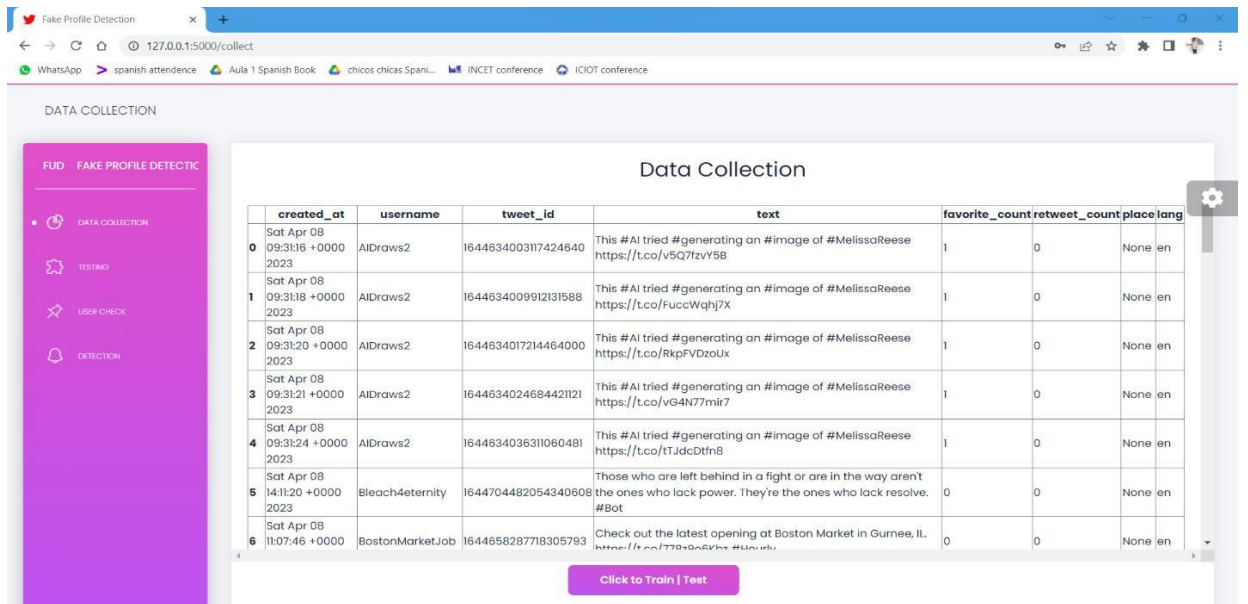
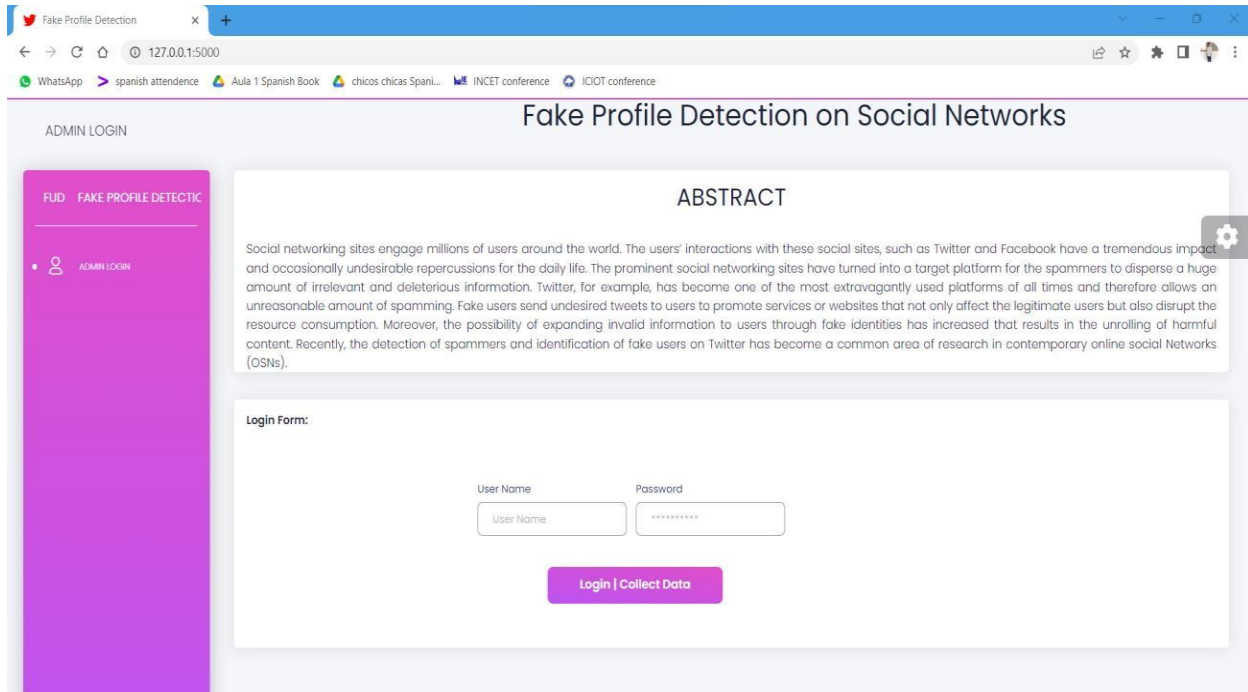
session['username'] = dl_user
download = download_user(dl_user)
df=pd.read_csv('coba.csv')
return render_template('check.html',df=df.to_html())

@app.route("/detect")
def detect():
    all_files = glob.glob(path + "/*.csv")
    for filename in all_files:
        if filename.endswith('.csv'):
            os.unlink(filename)
            #print(filename)
    my_var = session.get('username', None)
    print(my_var)
    df=pd.read_csv('coba.csv')
    prediksi,tab,y_pred=prediction(df)
    tab['Username']=my_var
    tab=tab.set_index('Username')
    tab=tab
    labels = 'Legitimate', 'Fake'
    size1 = 100-float(prediksi)
    size2 = prediksi
    sizes = [size1, size2]
    colors = ['forestgreen', 'crimson']
    explode = (0, 0.15)
    fig1, ax1 = plt.subplots()
    ax1.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%', colors =colors,
            shadow=True, startangle=90)
    ax1.axis('equal') # Equal aspect ratio ensures that pie is drawn as a circle.
    plt.savefig('static/images/graph/'+my_var+'.png',transparent=True)
    img = 'static/images/graph/'+my_var+'.png'
    return
    render_template('detect.html',prediction=prediksi,tab=tab.to_html(),y_pred=y_pred,my_
var=my_var,img=img)

if __name__ == "__main__":
    app.run(debug=True)

```

## B.SCREENSHOTS



Fake Profile Detection

127.0.0.1:5000/test

WhatsApp spanish attendance Aula 1 Spanish Book chicas chicas Spani... INCET conference ICIOT conference

### TEST RESULTS

FUD FAKE PROFILE DETECTIC

- DATA COLLECTION
- TESTING
- USER CHECK
- DETECTION

	id	id_str	Name	Username	Followers_count	Listed_count	Friends_count	Favorites_count	Created_at	
0	1621243240841183232	1621243240841183232	A.I. Gallery	AIDraws2	2	0	0	323	2023-02-02 20:24:51	F
1	1161255532814114817	1161255532814114817	Alpheus	alpheus_w	15	0	43	773	2019-08-13 12:37:48	F
2	1619798122795769856	1619798122795769856	Kevin	beverly0860	2	0	33	81	2023-01-29 20:42:38	F
3	268009026	268009026	朽ホルキア♥	Bleach4eternity	460	0	781	747	2011-03-18 00:09:23	F
4	818597386176069633	818597386176069633	Boston Market Jobs	BostonMarketJob	1074	23	126	7	2017-01-09 23:16:40	F
5	1062494193354911744	1062494193354911744	brickk**	bricklabss	133	5	491	4122	2018-11-13	F

Next | Check User

Fake Profile Detection

127.0.0.1:5000/check


WhatsApp spanish attendance Aula 1 Spanish Book chicas chicas Spani... INCET conference ICIOT conference

### USER CHECK

FUD FAKE PROFILE DETECTIC

- DATA COLLECTION
- TESTING
- USER CHECK
- DETECTION

### User Check

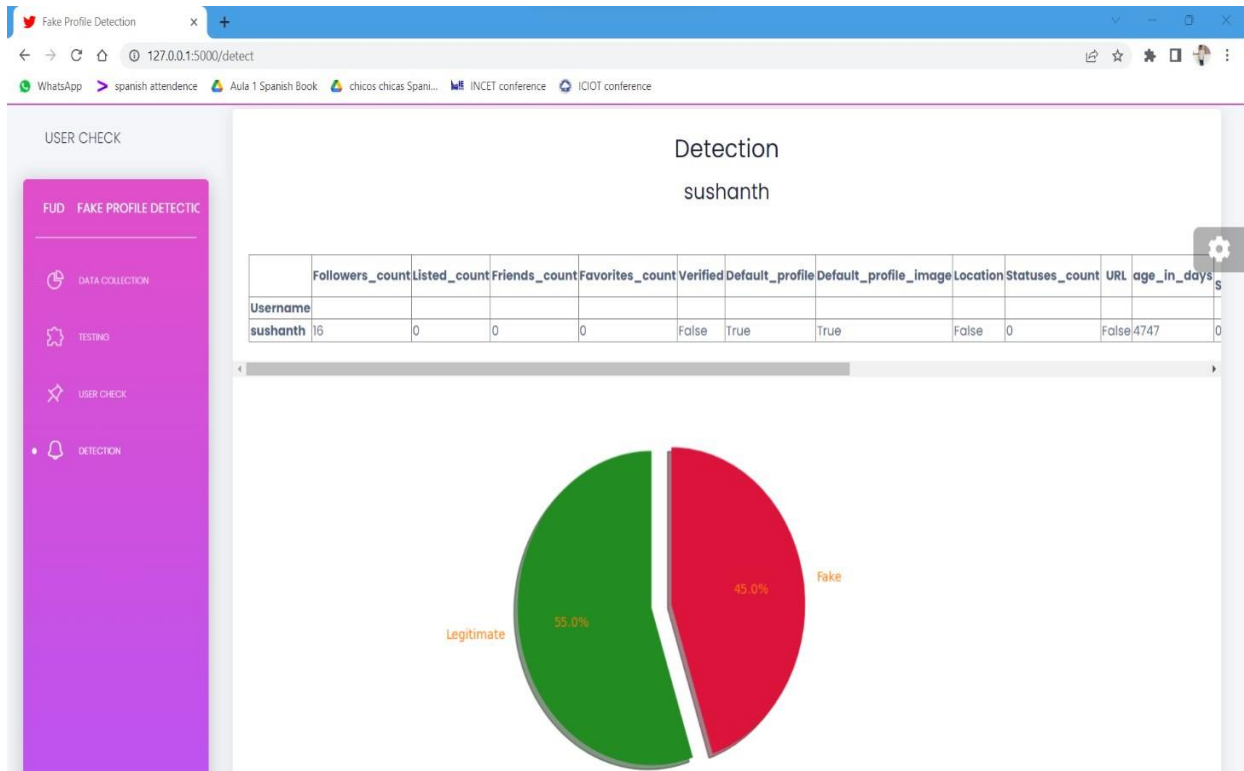


### Fake Profile Detection

Enter a username...

	id	id_str	Name	Username	Followers_count	Listed_count	Friends_count	Favorites_count	Created_at	Verified	Default_profile	Default_profile_image	
0	131119237131119237	131119237131119237	sushanth sundar	sushanth	16	0	0	0	2010-04-09 09:36:21	False	True	True	N

Next | Detect



## C.RESEARCH PAPER

# Email Phishing Messages Classification Using Machine Learning

Madallapalli Sushanth  
Dept of CSE  
Sathyabama University  
Chennai ,India  
msushanth55@gmail.com

C.A.Daphine Desona Clemency  
Dept of CSE  
Sathyabama University  
Chennai ,India  
daphine.cse@sathyabama.ac.in

Narreddy Murali Krishna Reddy  
Dept of CSE  
Sathyabama University  
Chennai ,India  
muralinarreddy008@gmail.com

### Abstract:-

*Vehicle reconciliation strategies have been involved a few times in spam channels to incorporate approaching/active messages, for example, spam and spam bunches. This technique expresses that each bunch contains little miniature groups, and each miniature group is dispersed. Notwithstanding, this thought ought not be trifled with, and the miniature group might have a lopsided dispersion. To build the respectability of the main strategy for appropriating the Internet class, we suggest supplanting the Euclidean space with a succession of models that incorporate into the miniature bunch connected with the circulation. Here, the Naïve Bayes (INB) classification has been carried out to carry out miniature bunches across the line. While these INBs can decide the distance and limits of micro clusters, Euclidean space considers the overall worth of the group and misdirects the bigger micro cluster. In this report, Den Stream is upheld by a committed framework called INB Den Stream. To represent the presentation of INB-Den Stream, current techniques, like Den Stream, StreamKM ++, and CluStream, have been utilized in the Twitter chronicle, and their exhibition has been estimated as far as quality, honesty as a general rule, memory as a rule, F1 measurements., markers, and complex estimations. The relatively close outcomes show that our techniques outflank its rivals in the set figures.*

**Keywords-** Spam Detection, Twitter, Car Training, Regular Forest, Certificate Tree, SVM.

### 1. INTRODUCTION

Web organizations (OSNs) like Facebook, Instagram, and Twitter have developed and become well known locales throughout the long term. As per the most recent figures, Twitter has 200 million individuals and gives in excess of 400 million tweets every day. A large number of these tweets incorporate spam, like publicizing messages, fishing assaults, dissemination of vindictive channels, and illegal tax avoidance. Spam tweets have normal elements, for example, "hashtags", "talk" and URLs in abridged terms,

however only one out of every odd message containing measures is spam. So this is a truly serious deal in the event that sifting spam works. Because of the approach of reallocation on Twitter, short URLs are frequently utilized by spammers to misdirect individuals. The capacity to conceal URLs is an intriguing objective to send, as Twitter doesn't have the foggiest idea where the URL is going . As well as concentrating on the substance of the message, you can screen the conduct of individuals to decide the wellspring of the spam. For instance, assuming a part's message surpasses the quantity of shared companions, all of their messages can be executed as spam. To hoodwink analysts, some message couriers who would rather not send modest quantities of spam or utilize counterfeit hashtags become a method for causing spam to show up in research. Because of the presence of spam measurements, AI strategies have been utilized to identify spam through informing in different OSNs. Until this point in time, controlled and uncontrolled strategies are being utilized, extended, and executed to guarantee that spam/spam is shipped off OSN. Obviously, the control strategy prompts preferable outcomes over the uncontrolled technique; However, the significant expense of introducing huge tokens diminishes the utilization of spam channels. Nonetheless, the consequences of the grouping technique show that the change to manage the thought isn't enough .

In this paper, we have fostered a better approach for voyaging that can uphold the customary approach to going instead of the Euclidean space and the development of the Naïve Bayes (INB) in the internet-based stage. In this article, we have picked DenStream in light of its exhibition (as far as link association) with work on its presentation. The interpreted rendition of INB's DenStream is classified "INB-DenStream". Since the local area is prepared to distinguish the middle and limits of the group on each miniature bunch that sounds more modern, these INB bundles

work with the effective dissemination of market-entering models. By checking the development of the miniature bunch populace over the long run, our program can change the design of the miniature group in an unmistakable and justifiable manner to change thoughts into message thoughts. Furthermore, the tweet-based conduct discussion was taken out from the Twitter file to deal with tweet-created content. The exhibition of our technique was contrasted with the present status of the bunch group as far as quality, respectability as a rule, memory as a rule, F1 estimation, and complex computations.

## II. LITERATURE REVIEW

F. Benevenuto, G. Magno, T. Rodriguez, and V. Almeida [2] contend that we approach the issue efficiently, however utilize a hashtag on Twitter to make preparing notes. Twitter is a well-known network that draws in numerous clients. A large number of these rundowns are not utilized because of spammers or occupation history needs or surveys. 89% of these worries are that clients have never made a spam estimation.

Z.Miller, B.Dickinson, W.Detrick, W.Hu, H.H. Wang on account data, designs, remarks, and client criticism reports.

C Divide spam and non-spam into two fundamental classifications, connecting test information to things that are hard for spammers to utilize and assist with recognizing spammers. Conventional spam channels don't function admirably on interpersonal organizations.

X. Zheng, Z. Zeng, Z. Twitter is a famous channel that has drawn in a great deal of clients.

A. Mukherjee et al. [6] The creators reasoned that the utilization of an enormous number of neurons made it challenging to acquire exact and sensible portrayals of preparing data progressively. A decent succession is addressed by a back-to-back tree or a little tree of the very shading that can be acquired in the normal request.

O.Kurasova, V.Martsinkevichus, V.Medvedev, A.Rapetska, P.Stefanovich nar. Numerous news sources stress that the degree of preparing assists them with learning better approaches to send spam and keep up with their insight into spam while looking for tweets. Erasing a spam client cannot channel spam messages, in light of the fact that the spammer can make another record and begin sending activities. An

identifier-based locator that registers tweets sent by confided in clients, contains no spam words, and shows the leftover highlights of the tweet.

A. H. Wang [11] claims that the classifier strategy is intended to deal with spam messages. The order cycle depends on a double worth calculation. The download work is a significant piece of the task to add advantages to the framework. 89% of spam accounts barely set up a client organization. The framework detailed non spam, how much data gave, and what the data gave meant to mental capacity.

G. Stringhini, C. Kruegel, and G. Vigna [12] disclosed that the capacity to spread unlawful data to clients through misleading data has upgraded the capacity to separate negative data. Counterfeit record clients' records are examined by the clients of the spam tweet account. It has been uncovered that many phony tweets are shared by supporters.

C. Yang, R. Harkreader, and G. Gu [14] tackle the issue, however use hashtags from Twitter to give preparing data. Twitter is a notable organization that draws in an enormous number of clients. We have tracked down that the capacity of the classifier to distinguish Twitter spam has lessened as it approaches reality.

## III. EXISTING SYSTEM

The traffic stream strategy has been involved ordinarily for spam examination to incorporate approaching/active messages, for example, spam and spam bunches. This technique expresses that each group contains little miniature bunches, and each miniature group is dispersed. Nonetheless, this thought ought not be messed with, and the miniature bunch might have a lopsided dispersion. To expand the honesty of the pre-web conveyance framework, we suggest supplanting the machine preparing machine. In light of our outcomes, the National Forest Service gives the best outcomes in the four areas we have assessed.

### DISADVANTAGES OF EXISTING SYSTEM

- Effective Strategies are not use.
- Real time records not used.

#### IV. PROPOSED SYSTEM

We talk about a portion of the various uses - in view of elements that separate among spammers and genuine clients. We promptly utilize this element to work with spam. Utilizing the Twitter-gave API strategy, we looked for dynamic Twitter clients, their supporters/data and the last 100 tweets. We then, at that point, looked into the agreement program in view of the gave information and the substance based. Carry out an AI calculation to make an assessment model. Then, at that point, we made a site utilizing jar. It will be executed as spam or non-spam. In view of our outcomes, the National Forest Service gives the best outcomes in the four areas we have assessed.

This study consist of a machine studying method proposed the use of the actual datasets & with numerous traits & development.

The proposed method is greater efficient & accurate than different present structure.

#### V. METHODOLOGY

##### MODULES:-

- Pre-handling of data
- Mechanical AI techniques

##### A. DATA PREPROCESSING

The first and most significant stage in quite a while handling is information assortment. We have assembled an information bundle in light of Twitter spam data. The informational collection and status data of the CSV document comprises of the quantity of Twitter spam messages. We should choose or eliminate highlights from the informational index we gather. Presently you really want to begin Data Cleaning. Consequently, the data handled in this module will be depleted.

##### B. MACHINE LEARNING MECHANISM

The informational index will be assessed utilizing four AI classifications: Tree Decision Support, Vector Classifier Support, Forest, and Naive Bayes Classifier Algorithm. Change to a calculation that contains spam and informational indexes in spam.

##### C. PERFORMANCE STATISTICS

The consequences of the task show that regular woods give the best outcomes in the four phases we have

thought of. Assessment should be possible as indicated by the accompanying models.

##### Execution plan

- The impact of various democratic strategies
- Convenient survey of data.

#### VI. IMPLEMENTATION

Here we gather genuine data on Twitter. From that point forward, we make a news program and read it. Since we have a spa data bundle, we will do this on the Machine Learning Algorithm. We utilize four sorts of calculations. We are using

- Tree declaration
- Support vector classifier
- Innocent Bayes classifier

##### A. CLASSIFICATION USING DECISION TREE

Picking a tree is a straightforward and simple to-utilize apparatus. When the tree is confirmed, it should be feasible to get to the patient line. The chose tree is separated into classes that are not difficult to manage and treachery. Choosing a tree should be possible in two stages by number. Measurements centers around the suspicion that data is useful and can be partitioned into three classifications.

Formula:-  $E(S) = -(P) \log_2 P - P(N) \log_2 P(N)$

Steps followed:-

Stage 1: Understand the significance of data related data.

Stage 2: Prepare the data in your heart to get the data in a decreased structure.

Stage 3: Give the best insight subsequent to learning the essentials of media.

Stage 4: Now utilize a comparative correlation with figure out the advantages of the data.

Stage 5: Reduce the level as per the expense of getting data.

Stage 6: Repeat the circle on each tree until everything transforms into a leaf.



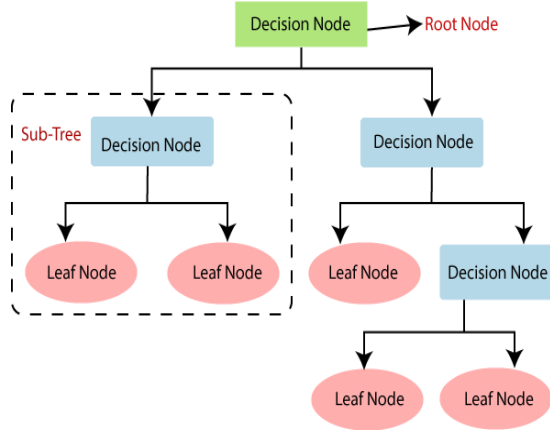


Fig-1

### B. CLASSIFICATION USING RANDOM FOREST

Standard memory is a calculation control machine that is generally utilized for arrangement and relapse issues. It constructs choice trees of various sizes and gets a large number of their arranging and looking at sounds as they pivot. It functions admirably for ordering issues. Long haul (RF) trees are a dependable logging blend, and each tree is free, separately, and in light of an interesting vector esteem chose exclusively. A typical issue with backwoods trees is the special strength of the wood and the interrelationships between them. It's significant regarding shouting. This is an arranged computation technique and is considered at the most elevated level on the grounds that the back tree is bigger than the handled tree. As a rule, the tree is planned autonomously and the tree is related with amicability. Customary memory numbers can be utilized to make hub arrangement issues.

Formula:-

$$RFf_i = \frac{\sum_j normf_{ij}}{\sum_{j \in \text{all features}, k \in \text{all trees}} normf_{jk}}$$

Steps followed:-

- Stage 1: The standard decision in different ways, here  $<< m$ .
- Stage 2: Using a decent security guide, count the middle "d" and circle the capacity.
- Stage 3: Minimize the young ladies line sharing point.
- Stage 4: Repeat stages 1-3 until the gatherings show up.
- Stage 5: Repeat the arrangement 1-4 times and return to the woods.

### C. NAIVE BAYES ALGORITHM

This is an arranging methodology in light of Base's hypothesis, which expects the autonomy of theories. So, Naive Bayes records his thought process is a sure element in classes that are not connected with different highlights. For instance, apples are a red natural product, around 3 crawls in breadth. Albeit these characteristics are connected with some attribute, these qualities assume a part in making it workable for these seeds to be apples, henceforth the name "credulous."

The Naive Bayes model is not difficult to construct and exceptionally helpful, particularly for enormous articles. Notwithstanding its straightforwardness, Innocent Bays is known for its predominance over the most grounded approach to positioning.

The base hypothesis gives a potential estimation of  $P(c | x)$  for  $P(c)$ ,  $P(x)$  and  $P(x | c)$ . See the accompanying delineation.

$$P(c | x) = \frac{P(x | c)P(c)}{P(x)}$$

Likelihood
Class Prior Probability

Posterior Probability
Predictor Prior Probability

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

### D. SUPPORT VECTOR MACHINE (SVM)

"Machine Support" (SVM) is the estimation of how much insight that can be utilized to isolate or turn. Be that as it may, it is utilized to recognize issues. In SVM estimations, we set every information thing as an article in the n-layer space (where n is the quantity of items you have), and the worth of each article is the worth of that regulator. We are right now concentrating on various hyperplanes to distinguish these two classes. Vector support is a blend of free cognizance works out. The SVM list is of two classes (hyper-airplane/line).

## VII. RESULTS

The calculations we analyze are Decision Tree Decision, Vector Classifier Support, Certified Forestry, and Naïve Bayes Algorithm. In stages, the classifications will give the best outcomes.

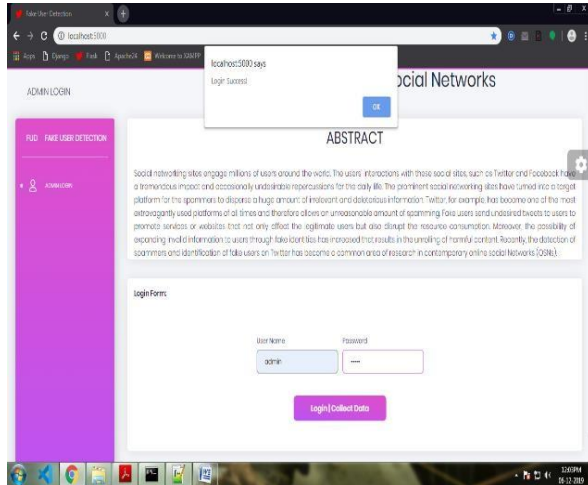


Fig-2

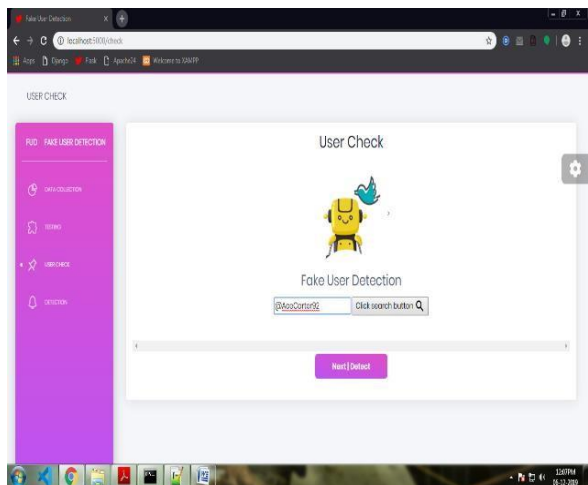


Fig-3

## CONCLUSION

In this , we proposed a deep studying model for emails unsolicited mail detection . We used UCI datasets for our project . We used 3 methods of words embedding, Being counted vectorized , subtract vectorize . And we used different algorithms .

## REFERENCES

- H. Tsukayama, Twitter Turns 7: Users Send Over 400 Million Tweets Per Day. Washington, DC, USA: Washington Post, Mar. 2013.
- F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in Proc. 7th Annu. Collaboration, Electron. Messaging, Anti Abuse Spam Conf., Redmond, WA, USA, Jul. 2010.
- Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Inf. Sci.*, vol. 260, pp. 64–73, Mar. 2014.
- C. Chen et al., "A performance evaluation of machine learning-based streaming spam tweets detection," *IEEE Trans. Comput. Social Syst.*, vol. 2, no. 3, pp. 65 76, Sep. 2015.
- X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing*, vol. 159, pp. 27–34, Jul. 2015.
- A. Mukherjee et al., "Spotting opinion spammers using behavioral footprints," in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Chicago, IL, USA, 2013, pp. 632–640.
- M. Taheri and R. Boostani, "Novel auxiliary techniques in clustering," in Proc. World Congr. Eng., London, U.K., 2007.
- H. Tajalizadeh and R. Boostani, "A Novel Clustering Framework for Stream Data Un nouveau cadre de classifications pour les données de flux," *Can. J. Elect. Comput. Eng.*, vol. 42, no. 1, pp. 27–33, 2018.
- F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in Proc. SIAM Int. Conf. Data Mining, 2006, pp. 59–70.
- A. H. Wang, "Don't follow me: Spam detection in twitter," in Proc. Int. Conf. Secur. Cryptogr. (SECRYPT), Jul. 2010, pp. 1–10.

