iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.60175

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

Email Platform: Spam Detection

Abhinav Kachole[1], Aniket Nagpure[2], Atharva Wagh[3], Prof. T. H. Patil[4]
Information Technology Department, Sinhgad College of Engineering, Pune, India

Abstract: ***In practically every industry today, from business to education, emails are used. Ham and spam are the two subcategories of emails. Email spam, often known as junk email or unwelcome email, is a kind of email that can be used to hurt any user by sapping their time and computing resources and stealing important data. Spam emailvolume is rising quickly day by day. Today's email and IoT service providers face huge and massive challenges with spam identification and filtration. Email filteringis one of the most important and well-known methods among all the methods createdfor identifying and preventing spam. SVM, decision trees, and other machine learning and deep learning approaches have all been applied to this problem. Together with the explosive growth in internet users, email spam has increased substantially in recent years. Individuals are using them for illegal and dishonest purposes, such as fraud, phishing, and distributing malicious links through unsolicited email that can harm our systems and attempt to access your systems. By quickly constructing phone-y/fake profiles and email accounts, spammers prey on those who are ignorant of these scams. They use a real name in their spam emails. As a result, it's critical to identify spam emails that include fraud. This project will accomplish this by utilizing machine learning methods, and this article will examine the machine learning algorithms, put them to use on our data sets, and select the approach that can detect emailspam with the maximum degree of precision and accuracy.***

I. INTRODUCTION

Email is the most used source of official communication method for business purposes. The usage of the email continuously increases despite of other methods of communications. Automated management of emails is important in the today's con-text as the volume of emails grows day by day. Out of the total emails, more than 55 percent is identified as spam. This shows that these spams consume email user time and resources generating no useful output. The spammers use developed and creative methods in order to fulfil their criminal activities using spam emails, There-fore, it is vital to understand different spam email classification techniques and their mechanism. Emails are widely used as a means of communication for personal and professional use. Detection of Phishing website is an intelligent and effective model that is based on using classification or association Data Mining algorithms. In the proposed system, detecting phished email can be described as a classification problem with two categories i.e. ham and phished. Machine Learning is a field of artificial intelligence in which the system is given the ability to learn without being explicitly programmed. In our model, supervised machine learning algorithms are used for classification.

II. AIMS &OBJECTIVE

1) Design and develop an approach for email phishing detection from large syn- thetic as well as real time data using machine learning.
2) To develop an approach using various machine learning algorithms and explore the accuracy using majority routing technique.
3) To develop algorithm for extract different kind of features from emails to achieve the better classification accuracy.
4) To validate and explore the system classification results with existing detection techniques.

III. PROBLEM DEFINITION

1) For the purpose of classification, multiple packet features were extracted from all emails in a self-made dataset which consists of n number of phished emails and m number of ham emails.
2) These features are fed into the classifiers and results noted. Aim is to use the least number of features to develop a system which provides higher accuracyand study the variation of features and classify using various machine learning algorithms.

IV. LITERATURE SURVEY

A. *Study Of Research Paper*

Paper Name 1: Email Spam Detection Using Machine Learning Algorithms
Author: Nikhil Kumar, Sanket Sonowal, Nishant

©IJRASET: All Rights are Reserved

Description: Email Spam has become a major problem nowadays, with Rapid growth of internet users, Email spams is also increasing. People are using them for illegal and unethical conducts, phishing and fraud. Sending malicious link through spam emails which can harm our system and can also seek in into your system.

Paper Name 2: Machine Learning Techniques for Spam Detection in Emailand IoT Platforms

Author: Naeem Ahmed, Rashid Amin, Hamza Aldabbas

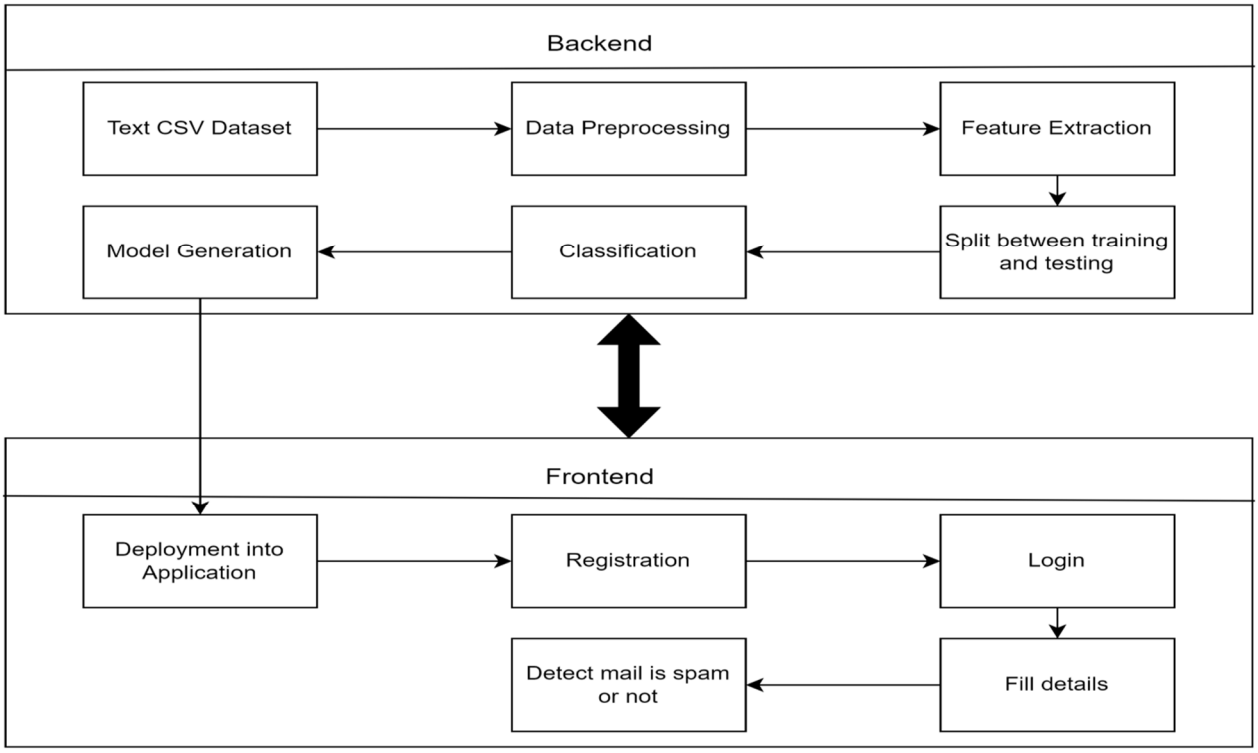
Description: Nowadays, emails are used in almost every field, from businessto education. Emails have two subcategories, i.e., ham and spam. Email spam,also called junk emails or unwanted emails, is a type of email that can be usedto harm any user by wasting his/her time, computing resources, and stealing valuable information.

Paper Name 3: Spam Detection Approach for Secure Mobile Message Communication Using Machine Learning Algorithms

Author: Luo Guang Jun, Shah Nazir, Habib Ullah Khan, Amin Ul Haq

Description: The spam detection is a big issue in mobile message communication due to which mobile message communication is insecure. In order to tackle this problem, an accurate and precise method is needed to detect the spam in mobile message communication. We proposed the applications of themachine learning-based spam detection method for accurate detection.

V. SYSTEM ARCHITECTURE



A. *Data Flow Diagram*

In Data Flow Diagram, we show that flow of data in our system in DFD0 in which rectangle present input as well as output and circle show our system. In DFD1 we show actual input and actual output of system input of our system is text or image and output is detected.

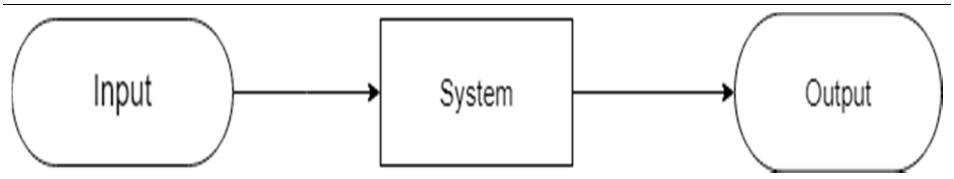


Figure: Data Flow diagram 0

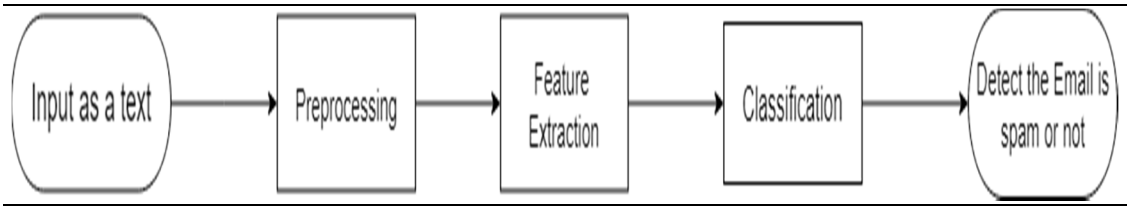
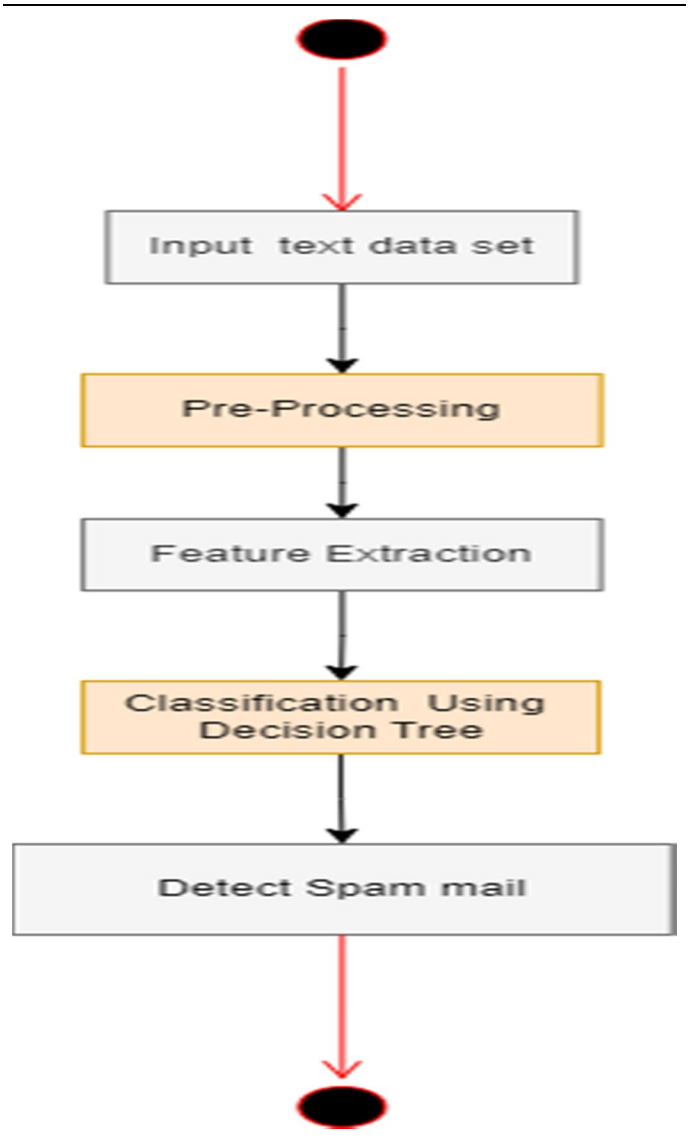


Figure: Data Flow diagram 1

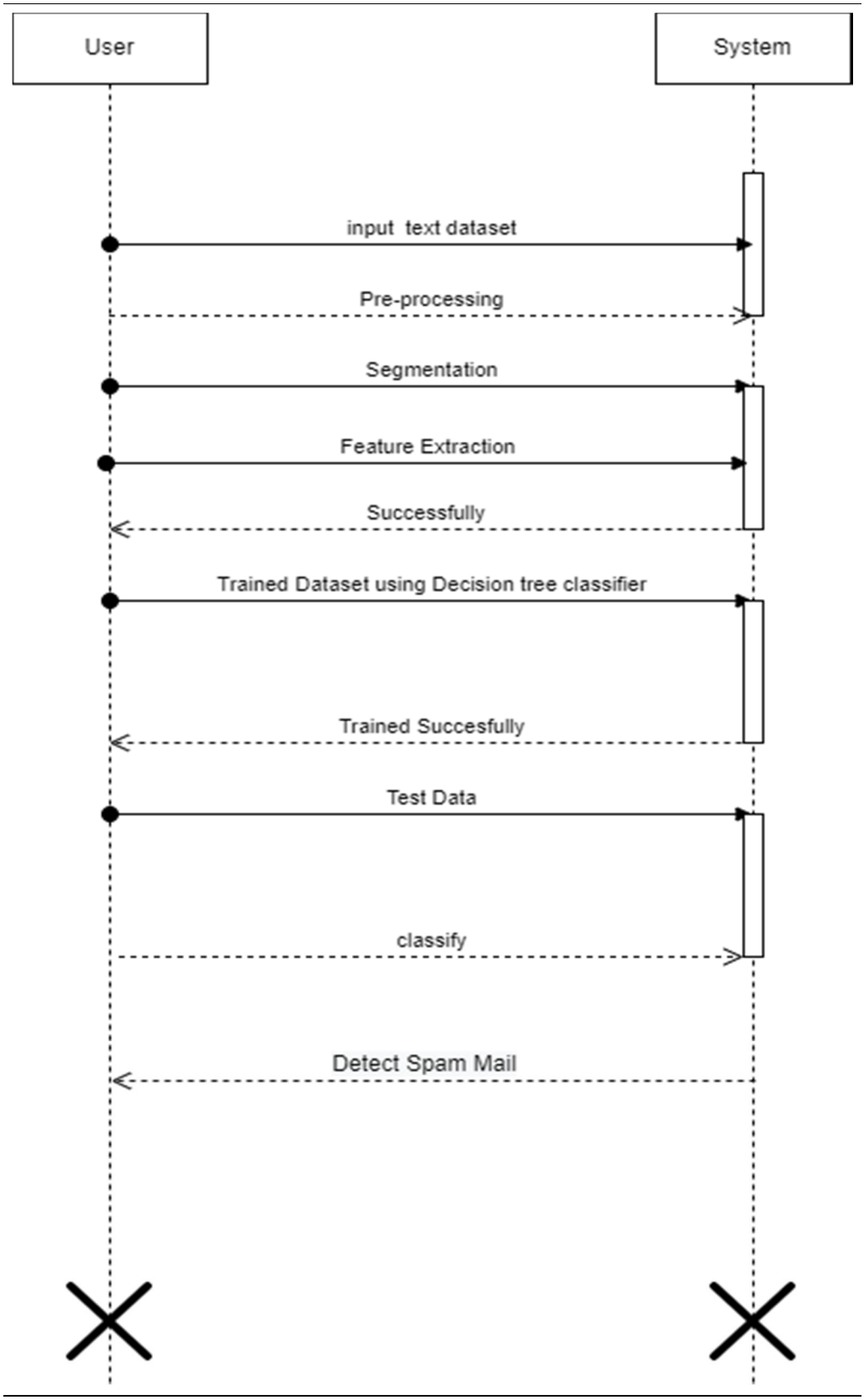
VI. UML DIAGRAMS

Unified Modeling Language is a standard language for writing software blueprints. UML diagrams may be used to visualize, specify, construct and document the artifacts of a software intensive system. UML is process independent, although optimally it should be used in process that is use case driven architecture-centric, iterative, and incremental. The Number of UML Diagram is available.

A. Activity Diagram

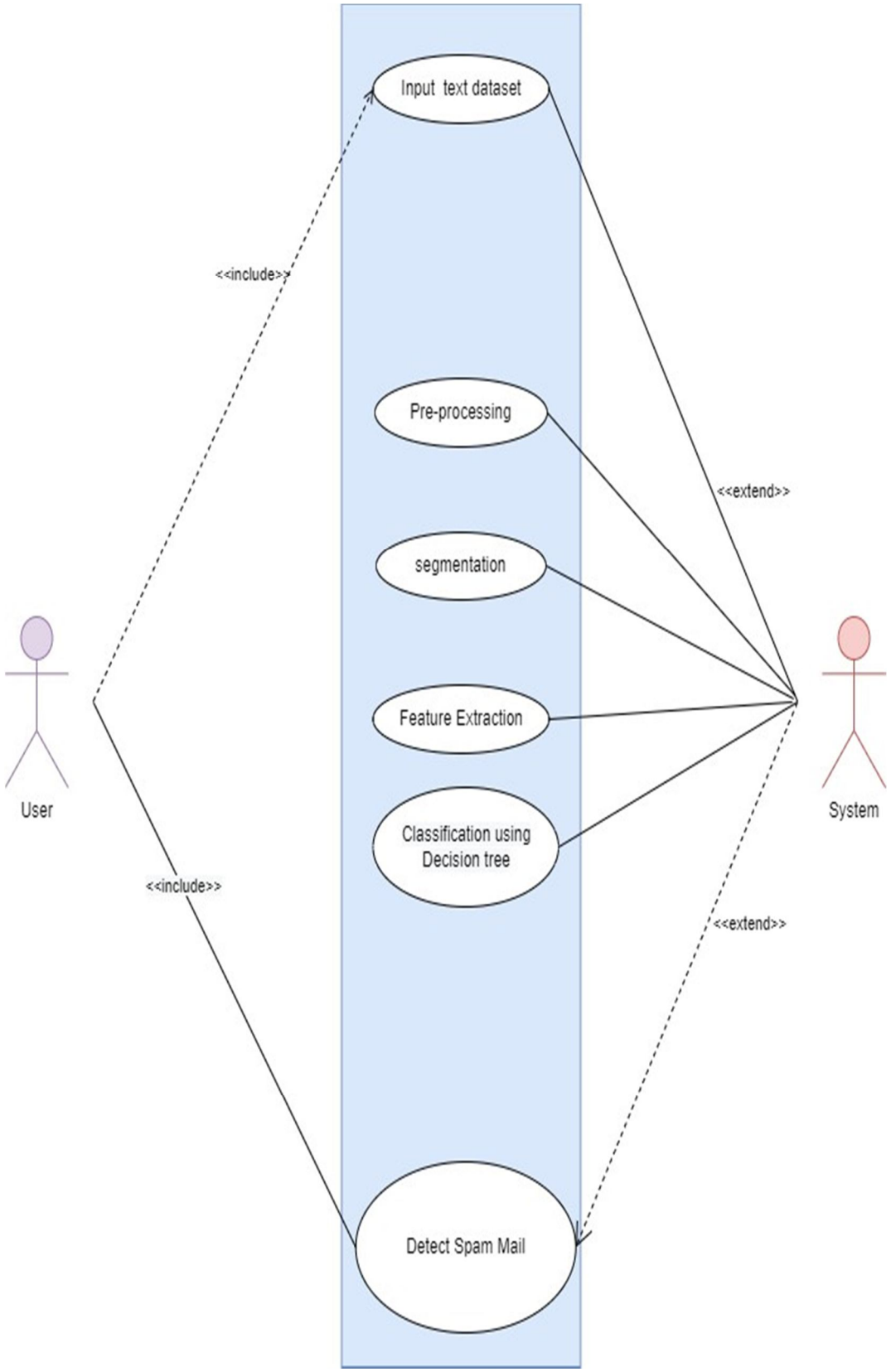


B. Sequence Diagram



©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. Use case Diagram



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

VII. PROJECT IMPLEMENTATION

A. Overview Of Project Modules

Pandas: Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series.This library is built on top of the NumPy library.

NumPy: NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

Import cv2: All packages contain Haar cascade files. cv2.data.haarcascades can be used as a shortcut to the data folder.

Pillow: Pillow is the friendly PIL fork by Alex Clark and Contributors. PILis the Python Imaging Library by Fredrik Lundh and Contributors.

B. Hardware Requirements

1) System Processors: Core2Duo
2) Speed: 2.4 GHz
3) Hard Disk: 150 GB

C. Software Requirements

1) Operating system: 32bit Windows 7 and on words
2) Coding Language: Python
3) IDE: Pycharm, Spyder
4) Database: Sqlite

D. Algorithms

Support Vector Machine: In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that as- signs new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

Two types of SVM

1) *Linear SVM:* Linear SVM is used for linearly separable data, which means ifa dataset can be classified into two classes by using a single straight line, thensuch data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
2) *Non-linear SVM:* Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non- linear SVM classifier.

Decision Tree: 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 nodesrepresent 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 solutionsto 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 split the tree into subtrees.

E. Advantages

1) Link base features extraction from mails.

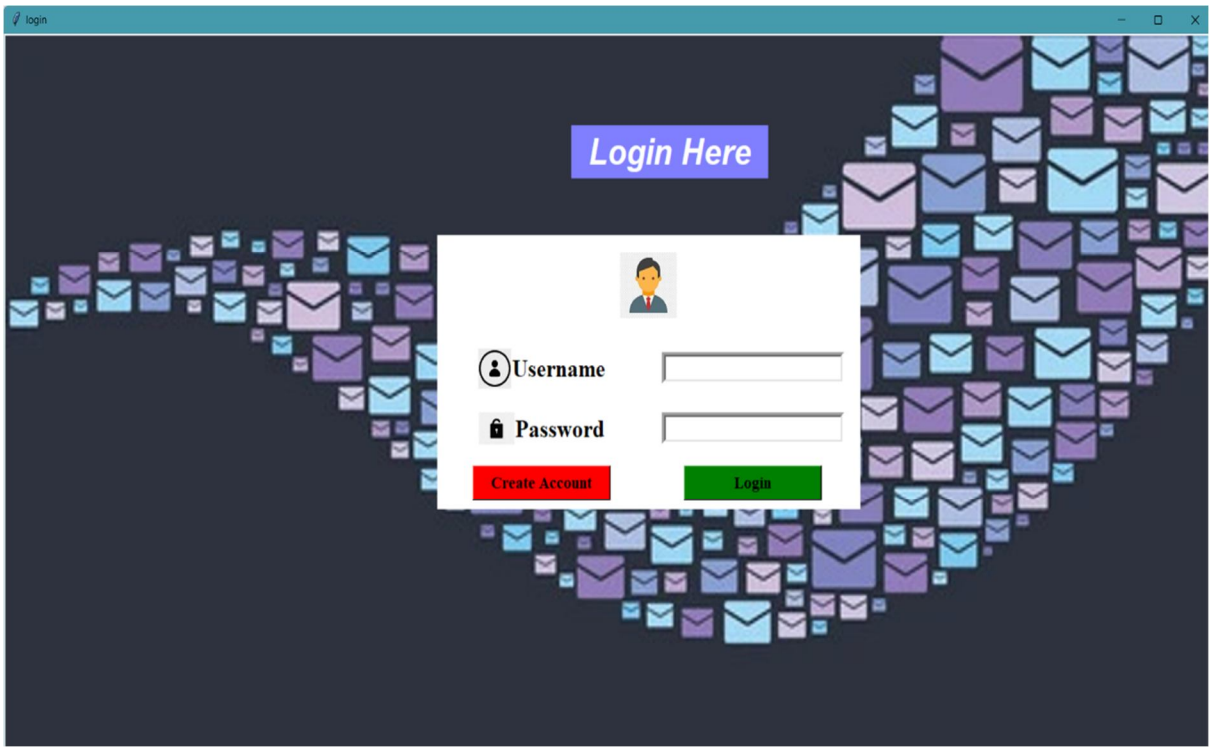
©IJRASET: All Rights are Reserved

2) Tag based features extraction for entire data.
3) Word base features extraction.
4) Classification result of all test data into fishing as well as normal respectively

VIII. EXPECTED OUTPUT

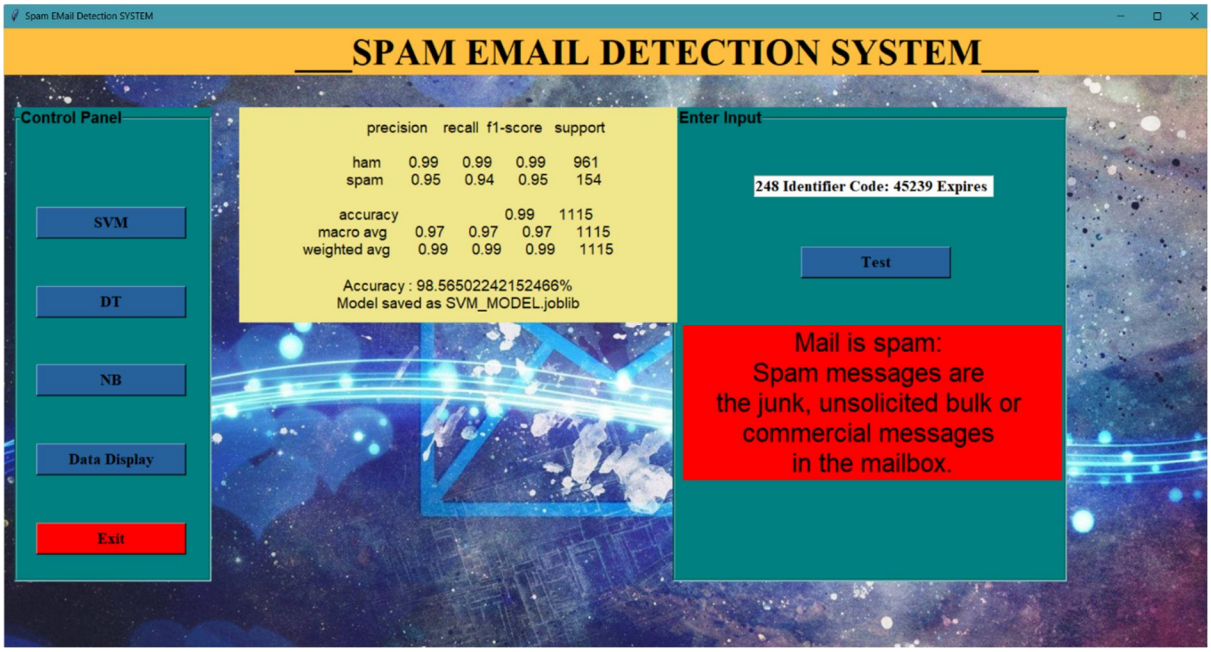
A. *Registration*

B. *Login*



©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. Application

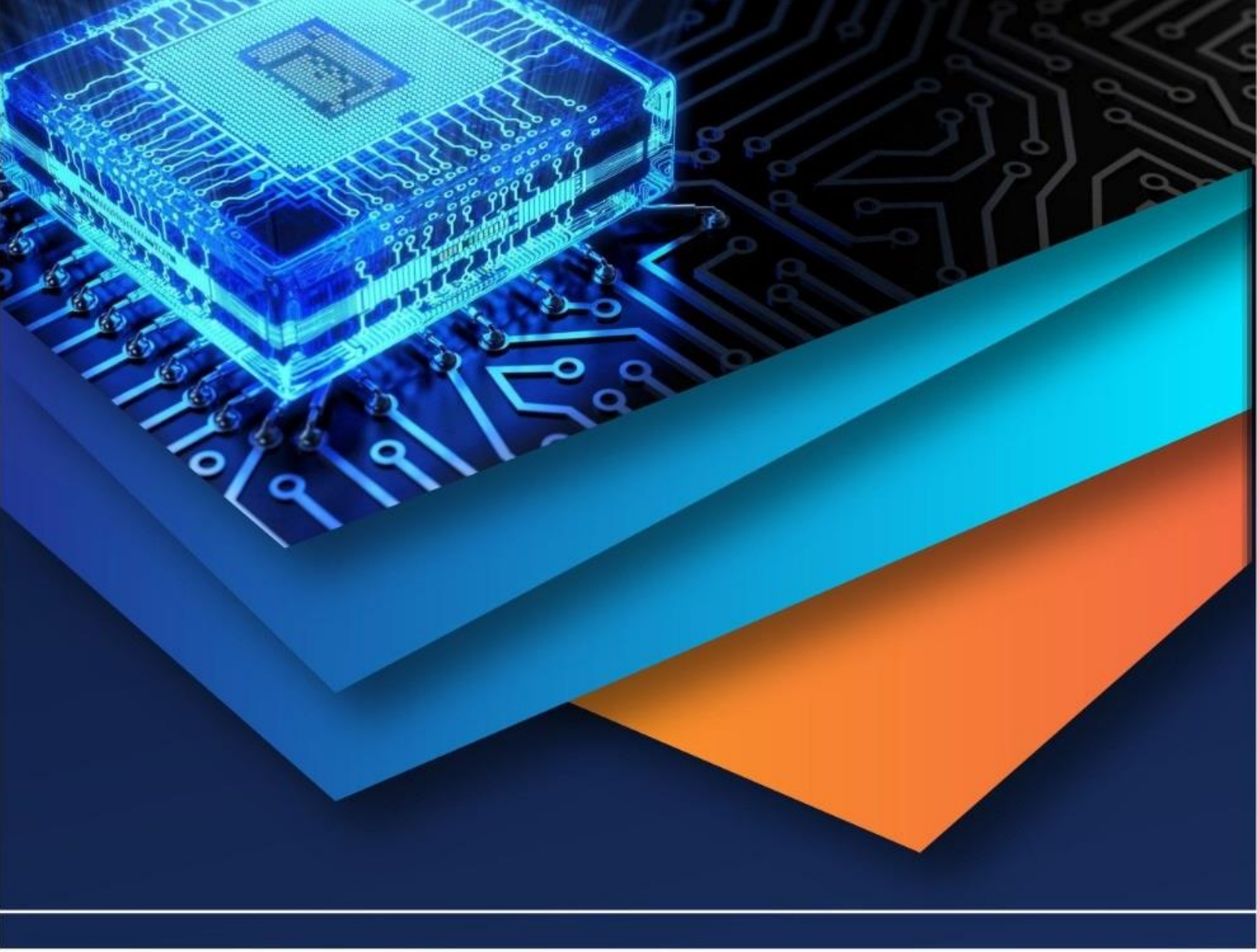


IX. CONCLUSION

Detection of spam is important for securing message and e-mail communication. The accurate detection of spam is a big issue, and many detection methods have been proposed by various researchers. However, these methods have a lack of capability to detect the spam accurately and efficiently. To solve this issue, we have proposed a method for spam detection using machine learning predictive models. Thus, the results suggest that the proposed method is more reliable for accurate and on-time detection of spam, and it will secure the communication systems of messages and e-mails. They conclude that most email spam filtering is done by utilizing Naïve Bayes and the SVM algorithm. To test the spam filtration models, these models can be trained on different datasets, such as "ECML" and UCI dataset.

REFERENCES

[1] Symantec, 2016 Internet Security Threat Report, 2016 (accessed May 20,2017).
[2] Kaspersky, Spam and phishing in Q3 2016, 2016 (accessed May 20, 2017).
[3] J. Alqatawna, A. Madain, A. M. Al-Zoubi, and R. AlSayyed, "Online social networks security: Threats, attacks, and future directions," in Social Media Shaping e-Publishing and Academia, pp. 121–132, Springer, 2017.
[4] J. Alqatawna, A. Hadi, M. Al-Zwairi, and M. Khader, "A preliminary analysis of drive-by email attacks in educational institutes," in Cybersecurity and Cyber forensics Conference (CCC), pp. 65–69, IEEE, 2016.
[5] R. A. Halaseh and J. Alqatawna, "Analyzing cybercrimes strategies: The case of phishing attack," in Cybersecurity and Cyber forensics Conference (CCC),pp. 82–88, IEEE, 2016.
[6] A. Zaid, J. Alqatawna, and A. Huneiti, "A proposed model for malicious spam detection in email systems of educational institutes," in Cybersecurity and Cyber forensics Conference (CCC), pp. 60–64, IEEE, 2016.
[7] A. M. Al-Zoubi, J. Alqatawna, and H. Faris, "Spam profile detection in social networks based on public features," in 2017 8th International Conference on Information and Communication Systems (ICICS), pp. 130–135, April 2017.
[8] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering," Expert Systems with Applications, vol. 36, no. 7,pp. 10206 – 10222, 2009.
[9] A. Rodan, H. Faris, and J. Alqatawna, "Optimizing feedforward neural net- works using biogeography based optimization for e-mail spam identification," International Journal of Communications, Network and System Sciences, vol.9, no. 1, p. 19, 2016.
[10] H. Faris, I. Aljarah, and J. Alqatawna, "Optimizing feedforward neural networks using krill herd algorithm for email spam detection," in AppliedElectrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on, vol, no, pp.1-5, pp. 1–5, IEEE, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)