A

Project Report on

Email Monitoring System

Submitted in partial fulfillment of the requirements

of the degree of

BACHELOR OF ENGINEERING

in

INFORMATION TECHNOLGY

(A.Y. 2022-2023)

by

Miss Asavari Pradeepkumar Mhadgut	(T-19-0028)
Miss Sayali Umesh Nalawade	(T-19-0024)
Miss Pratiksha Dilip Ranshur	(T-19-0200)

Under the Guidance of

Prof. Atiya Kazi

Assistant Professor, IT Department, FAMT

Department of Information Technology

Finolex Academy of Management and Technology, Ratnagiri

Year 2022 - 2023

CERTIFICATE

This is to certify that the Project Entitled “Email Monitoring System” is bonafied work of

Miss Asavari Pradeepkumar Mhadgut (T-19-0028)
Miss Sayali Umesh Nalawade (T-19-0024)
Miss Pratiksha Dilip Ranshur (T-19-0200)

Submitted to the University of Mumbai in partial fulfilment of the requirement for the award of the Degree of Bachelor of Engineering in Information Technology.

Signature: ____________________	Signature: ____________________
Name : Prof. Atiya Kazi	Name : Dr. Vinayak A. Bharadi
Assistant Professor	HOD-IT

Signature: ______________________________

Name : Dr. Kaushal Prasad
Principal,
Finolex Academy of Management and Technology.

Date:

Place: Finolex Academy of Management and Technology, Ratnagiri.

Project Report Approval for Bachelor of Engineering

This project report entitled ***Email Monitoring System*** by

Miss Asavari Pradeepkumar Mhadgut (T-19-0028)
Miss Sayali Umesh Nalawade (T-19-0024)
Miss Pratiksha Dilip Ranshur (T-19-0200)

is approved for the degree of ***BACHELOR OF ENGINEERING in INFORMATION TECHNOLOGY***.

Signature: ____________________

Name : Prof. Atiya Kazi
Assistant Professor

Signature: ____________________

Name : Dr. Vinayak A. Bharadi
HOD-IT

Signature: ______________________________

Name : Dr. Kaushal Prasad
Principal,
Finolex Academy of Management and Technology.

Examiners

1. Signature: ____________________
   Name :
2. Signature: ____________________
   Name :

Date:

Place:

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Signature)

1.---

2.---

3.---

(Name of student and Roll No.)

Date:

Index

Abstract

Nowadays, a big part of people relies on available email or messages sent by the stranger. The possibility that anybody can leave an email or a message provides a golden opportunity for spammers to write spam message about our different interests. Spam fills inbox with number of ridiculous emails. Degrades our internet speed to a great extent. Steals useful information like our details on our contact list. Identifying these spammers and also the spam content can be a hot topic of research and laborious tasks. Email spam is an operation to send messages in bulk by mail. Since the expense of the spam is borne mostly by the recipient, it is effectively postage due advertising. Spam email is a kind of commercial advertising which is economically viable because email could be a very cost effective medium for sender. Using Email Monitoring System. spam and non-spam emails are classified using ML algorithm like Naïve Bayes and KNN.

Keywords- Spam Emails, ML Algorithm, Email Monitoring

Chapter 1

Introduction

1.1. Introduction

Email has now become one of the best ways for advertisements due to which spam emails are generated. Spam emails are the emails that the receiver does not wish to receive. a large number of identical messages are sent to several recipients of email. Spam usually arises as a result of giving out our email address on an unauthorized or unscrupulous website. There are many of the effects of Spam. Fills our Inbox with number of ridiculous emails. Email Spam has become a major problem nowadays, with rapid growth of internet users. Spam email are the messages sent to multiple addresses. People are using them for illegal and unethical conducts such as phishing and frauds. So, it is needed to identify those spam mails which are fraud. Using Email Monitoring System, spam and non-spam emails are classified using ML algorithm like Naïve Bayes and KNN.

1.2. Literature Survey

We have reviewed some papers regarding our project “Email Spam Detection”.

In paper [1], authors have highlighted spam classification which is created using Bayes theorem and Naïve Bayes classifier and also IP addresses of the sender are often detected and its accuracy is 97%.

In paper [2], authors described cyber-attacks. Phishers and malicious attackers are frequently using email services to send false kinds of messages by which target user can lose their money and social reputations. It creates spam Email classification using Naïve Bayes Algorithm and the accuracy is 99%.

In paper [3], In this the system Uses Machine Learning techniques like Naïve Bayes, KNN, SVM and summarizes overall scenario regarding Accuracy Rate.

In paper [4], It compares implemented algorithm of XGboosting Classifier with existed classifier of SVM, Naïve Bayes and its accuracy is 95%.

In paper [5], It implements spam detection by using Machine learning and deep Learning Techniques like Naïve Bayes, Decision Tree, Neural Networks and random forest.

No.	Paper	Features
1	Email based spam detection	In this spam classification is created using Bayes theorem and Naïve Bayes classifier and also IP addresses of the sender are often detected. And its accuracy is 97% .
2	Spam email using Naïve Bayes Classifier	In this, it creates spam Email classification using Naïve Bayes Algorithm and the accuracy is 99%. Also calculates error rate
3	Existing Email Spam Filtering Method	In this the system Uses Machine Learning techniques like Naïve Bayes, KNN, SVM and summarizes overall scenario regarding Accuracy Rate.
4	Email spam filtering using ML	It compares implemented algorithm of XGboosting Classifier with existed classifier of SVM, Naïve Bayes and its accuracy is 95%
5	Machine Learning Techniques for spam Email	It implements spam detection by using Machine learning and deep Learning Techniques like Naïve Bayes, Decision Tree, Neural Networks and Random forest.

1.3. Motivation

We have done survey of few spam emails detection systems, from this we came to know that most of the people find the old traditional system which is not that much effective. Steals useful information like our details on our contact list. Identifying these spammers and also the spam content can be a hot topic of research and laborious tasks. So we saw this problem of youngsters or the people who use emails so oftenly. So this system, will help them to find which email is spam email and which is not.

1.4. Scope of the project

The system provides sensitivity to the clients and adapts well to the future spam techniques. Email monitoring system considers a complete message instead of single word with respect to its organization. It increases security and control. It reduces IT administration costs. In the future, we have a tendency to attempt to wear down tougher issues like the analysis and management of report in spam SMS filters storing. we are going to focus Transformer model for higher accuracy.

.

Chapter 2

Proposed Work

2.1. Problem Definition

Identifying these spammers and also the spam content can be a hot topic of research and laborious tasks. So, we saw this problem of youngsters or the people who use emails so oftenly. So, this system, will help them to find which email is spam email and which is not. The Existing Email system has a weak spam detection Mechanism. This results in segregation of important emails into spam emails. This can lead to miscommunication or delay of messages. Sometimes, some unimportant or phishing emails are not segregated into spam folder which leads unnecessary attention to unimportant emails. Drawbacks of existing system are as follows: It is source of viruses. It is capable to harm one’s computer and read out user’s e-mail address book and send themselves to number of people around the world. It can be source of various spams. These spam mails can fill up inbox and to deletion of these mail consumes lot of time. It is informal method of communication. The documents those require signatures are not managed by e-mail.

2.2. Requirement Analysis

1. Study of Existing System:

After going through many research paper and softwares or apps on the google play store and internet we found that existing system is not that much effective. There are some drawbacks of existing system. It is source of viruses. It is capable to harm one’s computer and read out user’s e-mail address book and send themselves to number of people around the world. It can be source of various spams. These spam mails can fill up inbox and to deletion of these mail consumes lot of time. It is informal method of communication. The documents those require signatures are not managed by e-mail.

2. Hardware Requirements :

Operating system : Windows

RAM : 8 GB(minimum requirement)

Hard Disk : 1GB working space (minimum requirement)

3. Software Requirements :

Languages : Python

Tools : Google Colab / Flask / Visual Studio

2.3 Proposed System

The main purpose of introducing this system is to help people who use emails very oftenly, to stay away from spam mails. These spam mails can fill up inbox and to deletion of these mail consumes lot of time. Sometimes, some unimportant or phishing emails are not segregated into spam folder which leads unnecessary attention to unimportant emails.

In this system to solve the problem of spam, email monitoring system is implemented using machine learning algorithms like Naive Bayes and KNN and accuracy of both these algorithms is compared.

Using this system, the problem of existing system is been solved.

2.4 Proposed Architecture

In our system to solve the problem of spam, email monitoring system is implemented using machine learning algorithms like Naive Bayes and KNN and accuracy of both these algorithms is compared. Using our system, the problem of existing system will be solved. There are some advantages of our system: Ensemble methods on the other hand proven to be useful as they using multiple classifiers for class prediction. Nowadays, lots of emails are sent and received and it is difficult as our project is only able to test emails using a limited amount of corpus. Our project, thus Email monitoring system is proficient of filtering mails giving to the content of the email and not according to the domain names or any other criteria. Also, our system has good efficiency as well as greater accuracy.

In our system, according to system architecture, incoming E-mails will be pre-processed and features are extracted and the features extracted are given to the training phase. And Email corpus is the text file of the email from which features are extracted and rough set rules are generated. Then using machine learning models like naïve bayes and KNN, emails are classified as spam and non-spam. The accuracy, precision and recall of naïve bayes and KNN will be compared. Flow chart, in which emails are collected from the dataset and it is pre-processed and filtered and the email will be detected as spam or non-spam. And finally, we will conclude that which algorithm is more accurate

2.5 System Architecture

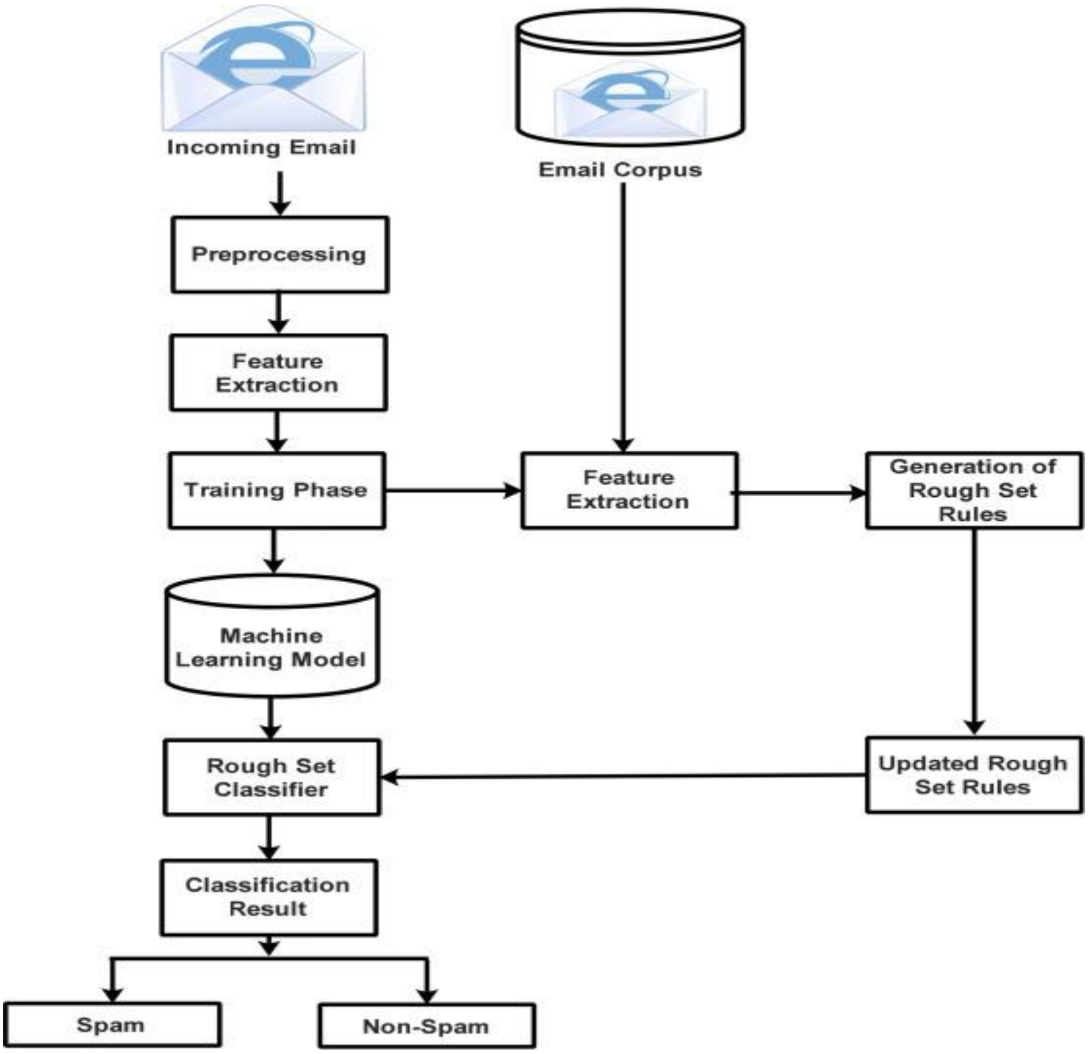


Fig 2. System Architecture for email monitoring system

In our system, according to Figure 2, incoming Emails will be pre-processed and features are extracted and the features extracted are given to the training phase. And Email corpus is the text file of the email from which features are extracted and rough set rules are generated. Then using the machine learning model - Logistic Regression, emails are classified as spam and non-spam.

2.6 Flow Diagram

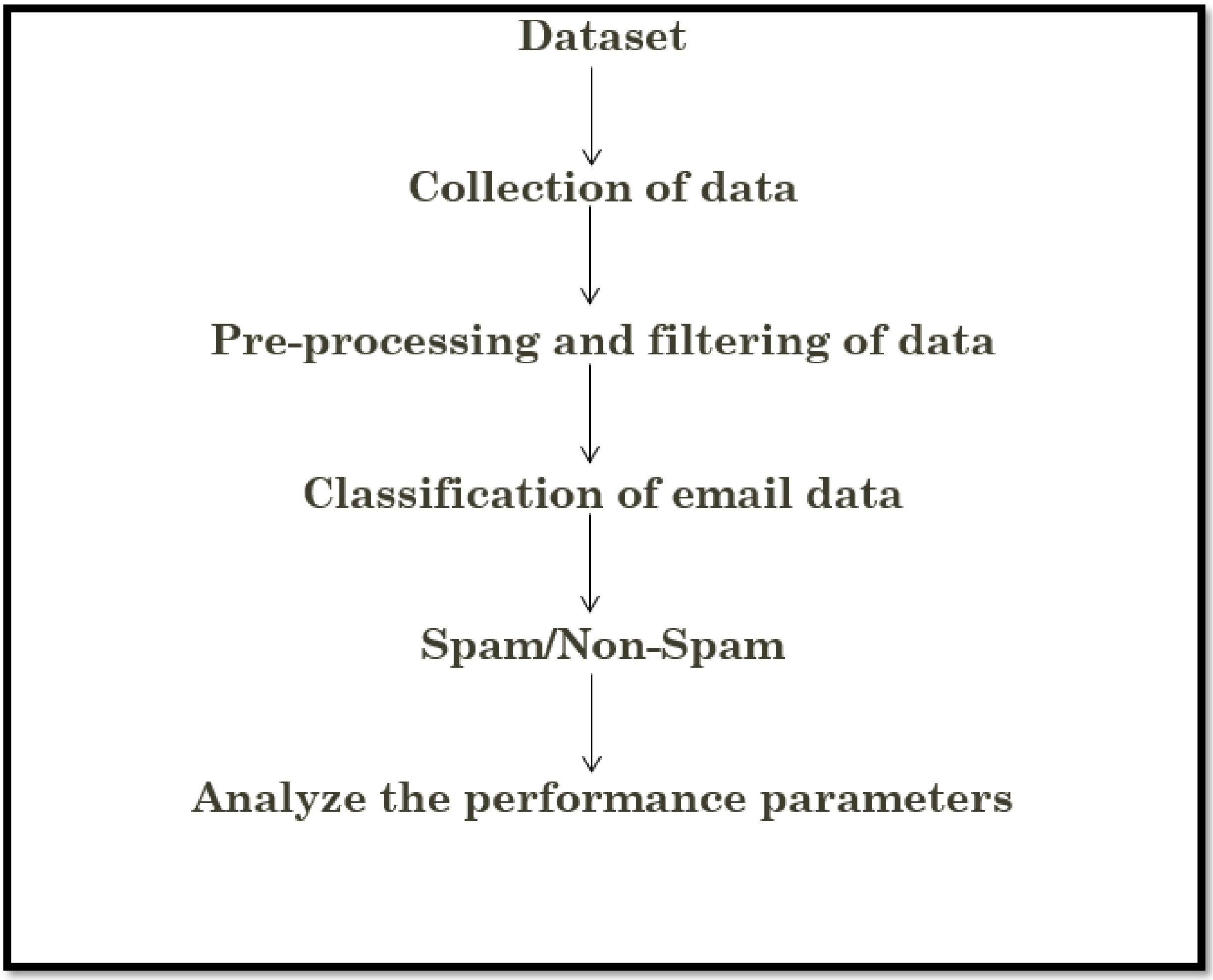

Dataset
Collection of data
Pre-processing and filtering of data
Classification of email data
Spam/Non-Spam
Analyze the performance parameters

Fig 1. Flow Diagram for email monitoring system

Figure 1 represents the flow chart of our system. In which emails are collected from the dataset and it is pre-processed and filtered and the email will be detected as spam or non-spam

2.7 Machine Learning Algorithm

What is Machine Learning?

A machine learning algorithm is the method by which the AI system conducts its task, generally predicting output values from given input data. The two main processes of machine learning algorithms are classification and regression.

Machine learning (ML) algorithms are broadly categorized as either supervised or unsupervised. Supervised learning algorithms have both input data and desired output data provided for them through labelling, while unsupervised algorithms work with data that is neither classified nor labelled. An unsupervised algorithm might, for example, group unsorted data according to similarities and differences.

However, many ML approaches, including transfer learning and active learning, involve what are more accurately described as semi-supervised algorithms. Transfer learning uses knowledge gained from completing one task to help solve a different but related problem, while active learning allows an algorithm to query the user or some other source for more information. Both systems are commonly used in situations where labelled data is scant.

Reinforcement learning, sometimes considered a fourth category, is based on rewarding desired behavioral and/or punishing undesired ones to direct unsupervised machine learning through rewards and penalties.

2.8 Algorithms used in the system

1. Naïve Bayes

2. KNN (K – Nearest Neighbor)

3. Logistic Regression

These are supervised ML algorithm used for classification problems.

Naïve Bayes is based on Bayes theorem which uses probability formula:

P(A|B) $= \frac{P(B|A)\,.\,P(A)}{P(B)}$

KNN is based on similarity which uses Euclidean distance formula:

$$\sqrt{(X1 - X2)^2 + (Y1 - Y2)^2}$$

Logistic regression represent by the following formula :

Logit(pi) = 1/(1+ exp(-pi))

ln(pi/(1-pi)) = Beta_0 + Beta_1*X_1 + … + B_k*K_k

2.9 Naïve Bayes

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in *text classification* that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

2.10 KNN

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

2.11 Logistic Regression

This type of statistical model (also known as *logit model*) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Logit(pi) = 1/(1+ exp(-pi))

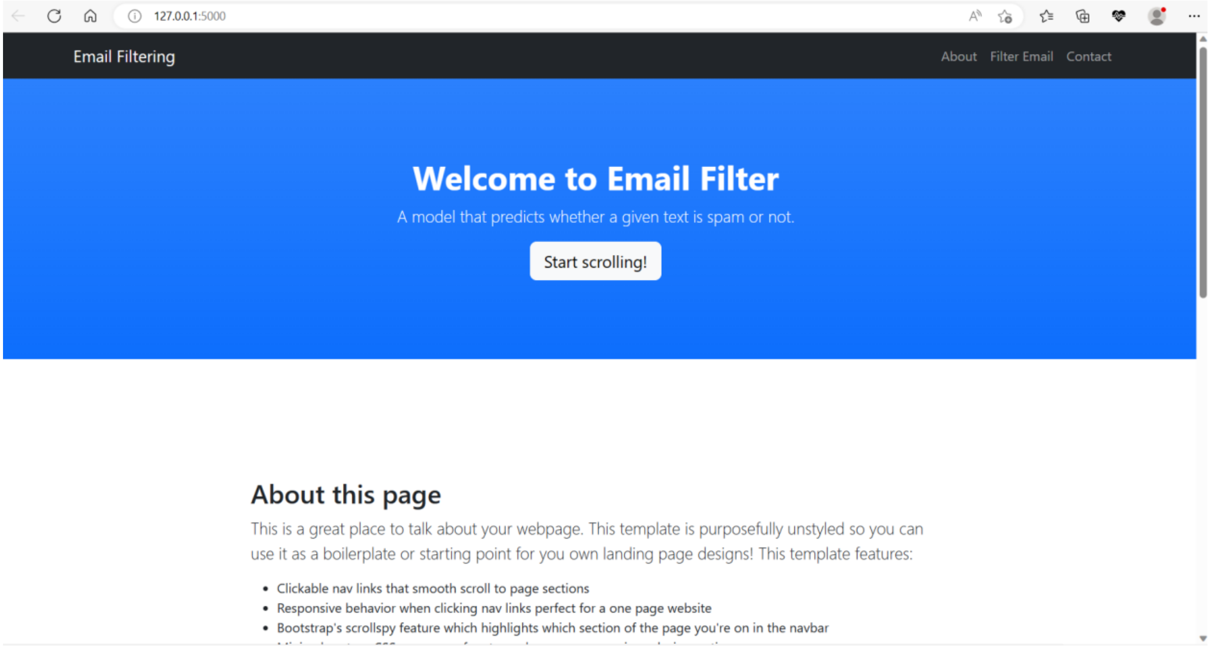
ln(pi/(1-pi)) = Beta_0 + Beta_1*X_1 + … + B_k*K_k

Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest

Chapter 3

Implementation Details

(Screenshots)



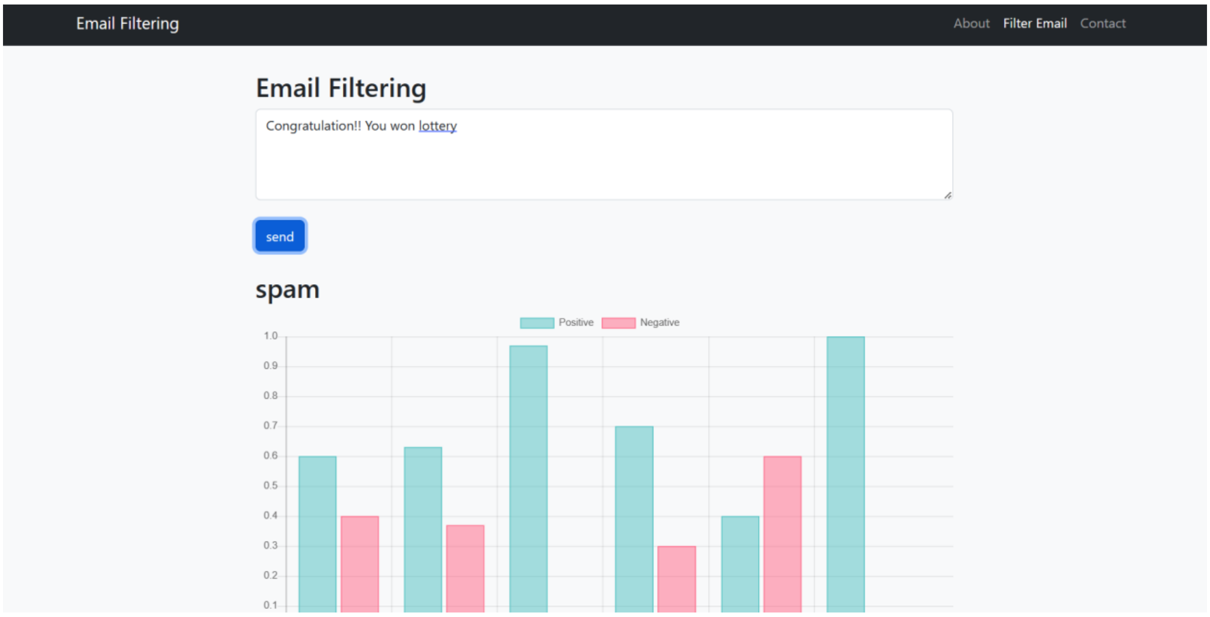
Welcome to Email Filter

A model that predicts whether a given text is spam or not.

About this page

This is a great place to talk about your webpage. This template is purposefully unstyled so you can use it as a boilerplate or starting point for you own landing page designs! This template features:

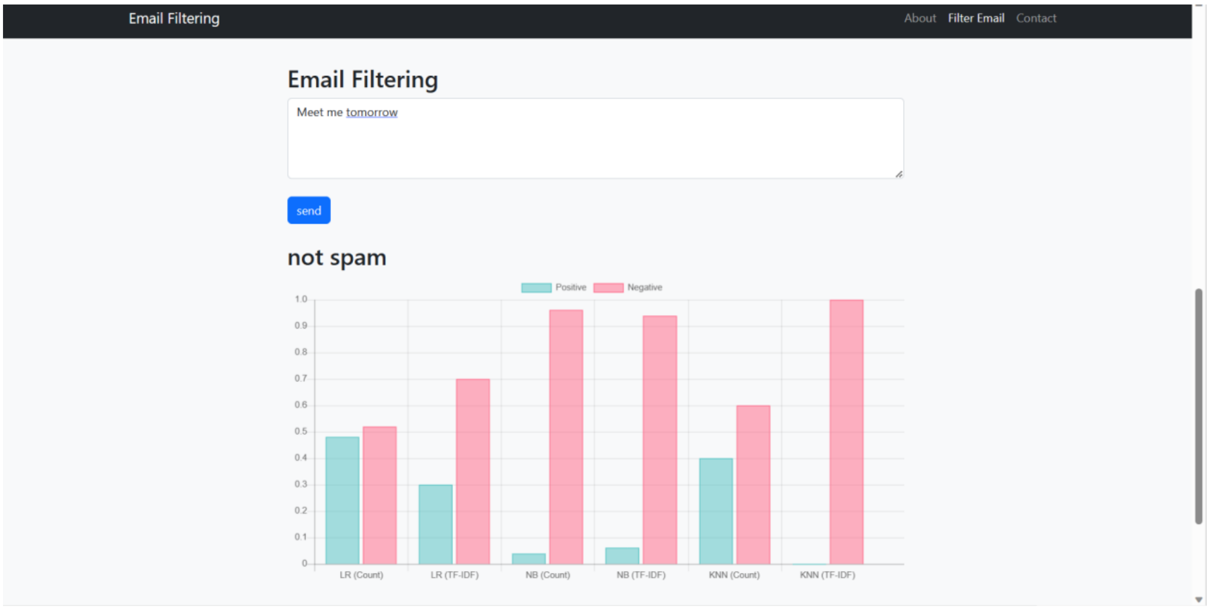
- Clickable nav links that smooth scroll to page sections
- Responsive behavior when clicking nav links perfect for a one page website
- Bootstrap's scrollspy feature which highlights which section of the page you're on in the navbar



Email Filtering

Congratulation!! You won lottery

spam



Email Filtering

Meet me tomorrow

not spam

Chapter 4

Analysis and Planning

4.1 Project Planning (Timeline)

Week No.	Planned Activity	Details of Implementation
1	Finalizing requirements	Planning all of the project requirements, including scope, quality, cost, schedule, etc.
2	Determine all roles & responsibilities & develop a schedule	Assigning roles and responsibilities to group members and scheduling start and end dates for each project activity.
3	Literature survey	Researching previous year research papers
4	Collecting resources and planning required technologies	Collection of all the resources required for the project and determining what technologies to use for the project
5	Designing UI/UX	Designing UI/UX using Flutter
6	Preparing and publishing journal	Referring to previous journals and creating and publishing a journal
7	Working on ML	Using a teachable machine to train data sets.
8	Working on Naïve Bayes	Using this algorithm to train data sets
9	Working on KNN	Using this algorithm to train data sets
10	analyzing and testing	analyzing and testing on mobile devices

4.2 Task Distribution

Task Name	Description	Start Date	End Date	Duration (Days)	Guide Remark
Task 1	Requirement gathering and analysis- Methods for determining requirements (Interview with interested parties, Brainstorming)	11-Jul-2022	15-Jul-2022	5	
Task 2	Literature Review	16-Jul-2022	31-Jul-2022	15	
Task 3	Software Requirement specification (Product Scope, Product perspective, product Functions, Operating Environment)	1-Aug2022	10-Aug2022	10	
Task 4	Feasibility Study Report	11-Aug2022	28-Aug2022	17	
Task 5	Project charter (Template)	1-Sep-2022	8-Sep-2022	8	
Task 6	Work-Breakdown Structure (Decomposition of the work)	9-Sep-2022	22-Sep-2022	13	
Task 7	Project scheduling- Gantt chart & Network Diagram	23-Sep2022	1-Oct-2022	9	
Task 8	Logical Design- UML diagram, Data flow Diagram	2-Oct-2022	12-Oct-2022	10	
Task 9	Analyzing the requirements for project	15-Oct2022	20-Oct-2022	5	
Task 10	Developing individual modules	21-Oct2022	30-Nov2022	40	
Task 11	Integrating all the modules	1-Dec-2022	4-Dec2022	4	
Task 12	Generating results for paper publication	5-Dec2022	7-Dec2022	3	
Task 13	Preparation of research paper	9-Jan2023	13-Jan2023	5	
Task 14	Testing of integrated modules	14-Jan2023	15-Feb-2023	33	
Task 15	Adapting the changes if any	16-Feb-2023	24-Mar-2023	37	
Task 16	Deployment of complete integrated system	25-Mar2023	7-Apr-2023	14	

4.3 Scheduling (Timeline Chart)

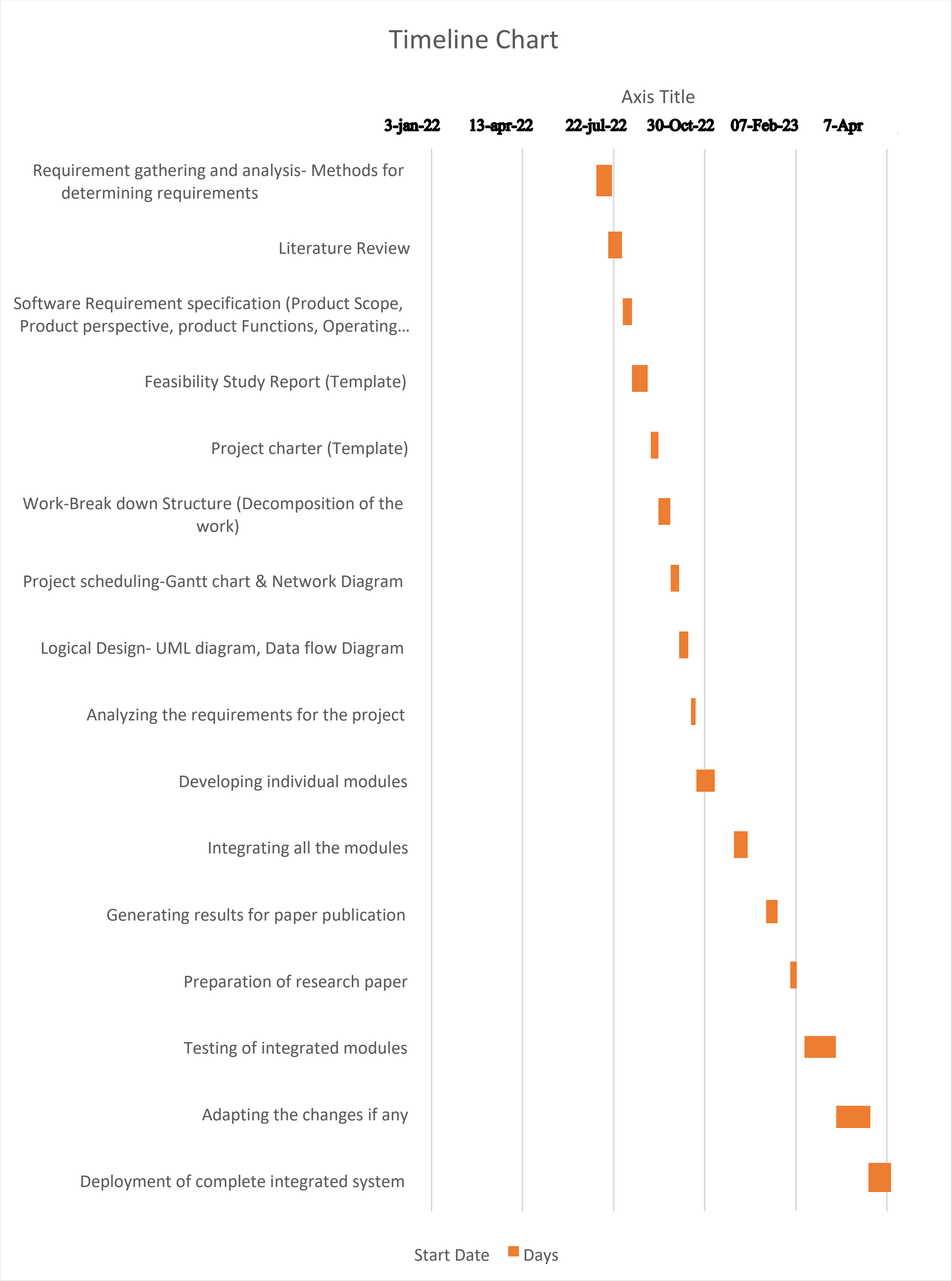


Figure : Timeline Chart

Chapter 5

Code

Main.py

```
import pandas as pd

import numpy as np

import joblib

from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

from sklearn.model_selection import train_test_split


df = pd.read_csv('final_random_spam_filtering_dataset.csv')

df['label'] = df.label.map({'ham':0,'spam':1})

X_train, X_test, y_train, y_test = train_test_split(df['text'], df['label'],
test_size=0.2, random_state=42)


# Vectorize the text data using CountVectorizer and TfidfVectorizer

vectorizer_count = CountVectorizer()

vectorizer_tfidf = TfidfVectorizer()

X_train_count = vectorizer_count.fit_transform(X_train)

X_train_tfidf = vectorizer_tfidf.fit_transform(X_train)


# Load the models for future use

loaded_knn_count_model = joblib.load('models/knn_count_model.joblib')

loaded_lr_tfidf_model = joblib.load('models/lr_tfidf_model.joblib')

loaded_nb_count_model = joblib.load('models/nb_count_model.joblib')

loaded_knn_tfidf_model = joblib.load('models/knn_tfidf_model.joblib')

loaded_lr_count_model = joblib.load('models/lr_count_model.joblib')

loaded_nb_tfidf_model = joblib.load('models/nb_tfidf_model.joblib')

new_text = ["Reminder: Our team meeting is scheduled for tomorrow at 10am in
the conference room."]
```

```
new_text_count = vectorizer_count.transform(new_text)
new_text_tfidf = vectorizer_tfidf.transform(new_text)
# Use the trained models to predict whether the new text is spam or ham
lr_count_prediction = loaded_lr_count_model.predict(new_text_count)
lr_tfidf_prediction = loaded_lr_tfidf_model.predict(new_text_tfidf)
nb_count_prediction = loaded_nb_count_model.predict(new_text_count)
nb_tfidf_prediction = loaded_nb_tfidf_model.predict(new_text_tfidf)
knn_count_prediction = loaded_knn_count_model.predict(new_text_count)
knn_tfidf_prediction = loaded_knn_tfidf_model.predict(new_text_tfidf)
lr_count_scores = loaded_lr_count_model.predict_proba(new_text_count)
lr_tfidf_scores = loaded_lr_tfidf_model.predict_proba(new_text_tfidf)

nb_count_scores = loaded_nb_count_model.predict_proba(new_text_count)
nb_tfidf_scores = loaded_nb_tfidf_model.predict_proba(new_text_tfidf)

knn_count_scores = loaded_knn_count_model.predict_proba(new_text_count)
knn_tfidf_scores = loaded_knn_tfidf_model.predict_proba(new_text_tfidf)

print('lr_count_prediction',lr_count_prediction)
print('lr_count_scores:', lr_count_scores)
print('lr_tfidf_prediction',lr_tfidf_prediction)
print('lr_tfidf_scores:', lr_tfidf_scores)
print('nb_count_prediction',nb_count_prediction)
print('nb_count_scores:', nb_count_scores)
print('nb_tfidf_prediction',nb_tfidf_prediction)
print('nb_tfidf_scores:', nb_tfidf_scores)
print('knn_count_prediction',knn_count_prediction)
print('knn_count_scores:', knn_count_scores)
print('knn_tfidf_prediction',knn_tfidf_prediction)
print('knn_tfidf_scores:', knn_tfidf_scores)
zeros_prob = []
ones_prob = []
lr_count_zero,lr_count_one = np.split(lr_count_scores[0],2)
lr_tfidf_zero,lr_tfidf_one = np.split(lr_tfidf_scores[0],2)
nb_count_zero,nb_count_one = np.split(nb_count_scores[0],2)
nb_tfidf_zero,nb_tfidf_one = np.split(nb_tfidf_scores[0],2)
knn_count_zero,knn_count_one = np.split(knn_count_scores[0],2)
knn_tfidf_zero,knn_tfidf_one = np.split(knn_tfidf_scores[0],2)
zeros_prob.extend([lr_count_zero,lr_tfidf_zero,nb_count_zero,nb_tfidf_zero,knn_count_zero,knn_tfidf_zero])
ones_prob.extend([lr_count_one,lr_tfidf_one,nb_count_one,nb_tfidf_one,knn_count_one,knn_tfidf_one])
zeros_prob = [round(score.tolist()[0],2) for score in zeros_prob]
ones_prob = [round(label.tolist()[0],2) for label in ones_prob]

accumulate_result = []
accumulate_result.extend([lr_count_prediction,lr_tfidf_prediction,nb_count_prediction,nb_tfidf_prediction,knn_count_prediction,knn_tfidf_prediction])
accumulate_result = [round(label.tolist()[0],2) for label in accumulate_result]
final_result = 'not spam' if accumulate_result.count(0) > accumulate_result.count(1) else 'spam'
print(zeros_prob,ones_prob, accumulate_result,final_result)
```

Chapter 6

Conclusion

Conclusion and future work

Email has been the most important medium of communication nowadays; through internet connectivity any message can be delivered to all aver the world. More than 270 billion emails are exchanged daily, about 57% of these are just spam emails. Spam emails, also known as non-self, are undesired commercial or malicious emails, which affects or hacks personal information like bank, related to money or anything that causes destruction to single individual or a corporation or a group of people. Hence this system is designed in such a way that it detects unsolicited and unwanted emails and prevents them hence helping in reducing the spam message which would be of great benefit to individuals as well as to the company. In the future this system can be implemented by using different algorithms and also more features can be added to the existing system. We will implement email monitoring system using Machine Learning Algorithms such as Naive Bayes Algorithm and KNN (K-Nearest Neighbor) Algorithm and compare which is better

Chapter 7

References

References

Journal Paper,

[1] “Email based Spam Detection” Thashina Sultana, K A Sapnaz, Fathima Sana, Mrs. Jamedar Najath, Dept. of Computer Science and Engineering, Yenepoya Institute of Technology Moodbidri, India

[2] “Email Spam Detection using Naive Bayes Classifier” Megha Tope, ME Student, Computer Science and Engineering, CSMSS College of Engineering, Aurangabad, India

[3] “A Survey of Existing E-Mail Spam Filtering Methods” Considering Machine Learning Techniques By Hanif Bhuiyan, Akm Ashiquzzaman, Tamanna Islam Juthi, Suzit Biswas & Jinat Ara, Southeast University

[4] “Email Spam Filtering Using Machine Learning Based Xgboost Classifier Method”, 1p.U. Anitha, 2dr.C.V. Guru Rao, 3dr. D. Suresh Babu, 1PhD Scholar, Dept of Computer Science and Engineering, JNTU, Hyderabad 2DIRECTOR, S.R. Engineering College, Warangal, Telangana, India, 3H.O.D, Department of Computer Science and Engineering, Kakatiya Government College.

[5] “Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges”, Naeem Ahmed, Rashid Amin, Hamza Aldabbas, Deepika Koundal, Bader Alouffi, and Tariq Shah.

[6] “Identification of Spam Email Based on Information from Email Header”, Shukor Bin Abd Razak, Ahmad Fahrulrazie Bin Mohamad 13th International Conference on Intelligent Systems Design and Applications (ISDA), 2013.

[7] “E-Mail Spam Detection Based on Part of Speech Tagging”, Mohammed Reza Parsei, Mohammed Salehi 2nd International Conference on Knowledge Based Engineering and Innovation (KBEI), 2015.

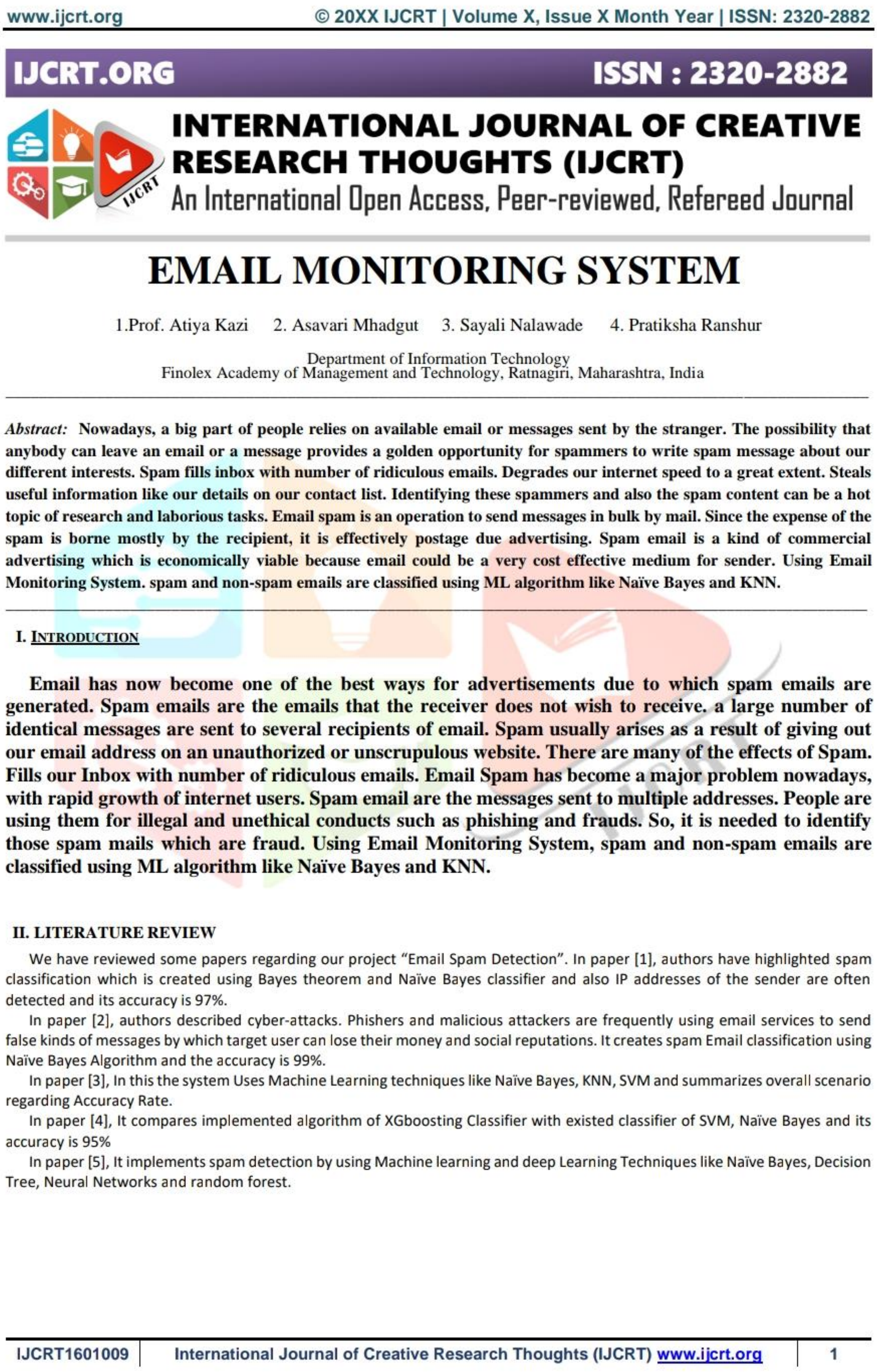
[8] "Email classification research trends: Review and open issues",Mujtaba, Ghulam, et al. IEEE Access 5 (2017).

[9] "Spam/ham e-mail classification using machine learning methods based on bag of words technique", Sahın, Esra, Murat Aydos, and Fatih Orhan, 2018 26th Signal Processing and Communications ApplicationsConference (SIU). IEEE, 2018.

Chapter 8

Publications

8.1 Conference Paper



IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EMAIL MONITORING SYSTEM

1.Prof. Atiya Kazi 2. Asavari Mhadgut 3. Sayali Nalawade 4. Pratiksha Ranshur

Department of Information Technology
Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India

Abstract: **Nowadays, a big part of people relies on available email or messages sent by the stranger. The possibility that anybody can leave an email or a message provides a golden opportunity for spammers to write spam message about our different interests. Spam fills inbox with number of ridiculous emails. Degrades our internet speed to a great extent. Steals useful information like our details on our contact list. Identifying these spammers and also the spam content can be a hot topic of research and laborious tasks. Email spam is an operation to send messages in bulk by mail. Since the expense of the spam is borne mostly by the recipient, it is effectively postage due advertising. Spam email is a kind of commercial advertising which is economically viable because email could be a very cost effective medium for sender. Using Email Monitoring System. spam and non-spam emails are classified using ML algorithm like Naïve Bayes and KNN.**

I. INTRODUCTION

Email has now become one of the best ways for advertisements due to which spam emails are generated. Spam emails are the emails that the receiver does not wish to receive. a large number of identical messages are sent to several recipients of email. Spam usually arises as a result of giving out our email address on an unauthorized or unscrupulous website. There are many of the effects of Spam. Fills our Inbox with number of ridiculous emails. Email Spam has become a major problem nowadays, with rapid growth of internet users. Spam email are the messages sent to multiple addresses. People are using them for illegal and unethical conducts such as phishing and frauds. So, it is needed to identify those spam mails which are fraud. Using Email Monitoring System, spam and non-spam emails are classified using ML algorithm like Naïve Bayes and KNN.

II. LITERATURE REVIEW

We have reviewed some papers regarding our project "Email Spam Detection". In paper [1], authors have highlighted spam classification which is created using Bayes theorem and Naïve Bayes classifier and also IP addresses of the sender are often detected and its accuracy is 97%.

In paper [2], authors described cyber-attacks. Phishers and malicious attackers are frequently using email services to send false kinds of messages by which target user can lose their money and social reputations. It creates spam Email classification using Naïve Bayes Algorithm and the accuracy is 99%.

In paper [3], In this the system Uses Machine Learning techniques like Naïve Bayes, KNN, SVM and summarizes overall scenario regarding Accuracy Rate.

In paper [4], It compares implemented algorithm of XGboosting Classifier with existed classifier of SVM, Naïve Bayes and its accuracy is 95%

In paper [5], It implements spam detection by using Machine learning and deep Learning Techniques like Naïve Bayes, Decision Tree, Neural Networks and random forest.

III. EXISTING SYSTEM

The Existing Email system has a weak spam detection Mechanism. This results in segregation of important emails into spam emails. This can lead to miscommunication or delay of messages. Sometimes, some unimportant or phishing emails are not segregated into spam folder which leads unnecessary attention to unimportant emails.

Drawbacks of existing system are as follows: It is source of viruses. It is capable to harm one's computer and read out user's e-mail address book and send themselves to number of people around the world. It can be source of various spams. These spam mails can fill up inbox and to deletion of these mail consumes lot of time. It is informal method of communication. The documents those require signatures are not managed by e-mail.

IV. PROPOSED ARCHITECTURE

In our system to solve the problem of spam, email monitoring system is implemented using machine learning algorithms like Naive Bayes and KNN and accuracy of both these algorithms is compared. Using our system, the problem of existing system will be solved.

There are some advantages of our system: Ensemble methods on the other hand proven to be useful as they using multiple classifiers for class prediction. Nowadays, lots of emails are sent and received and it is difficult as our project is only able to test emails using a limited amount of corpus. Our project, thus Email monitoring system is proficient of filtering mails giving to the content of the email and not according to the domain names or any other criteria. Also, our system has good efficiency as well as greater accuracy.

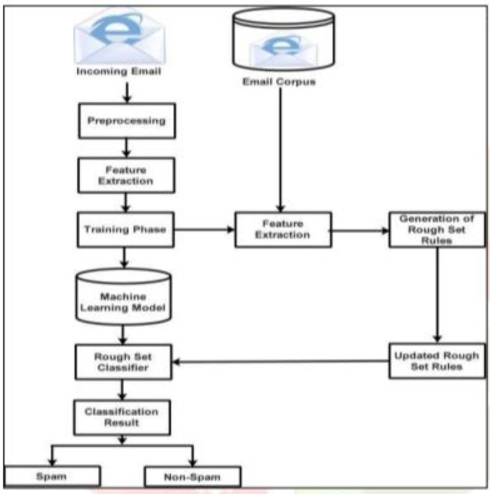


Fig1. System Architecture

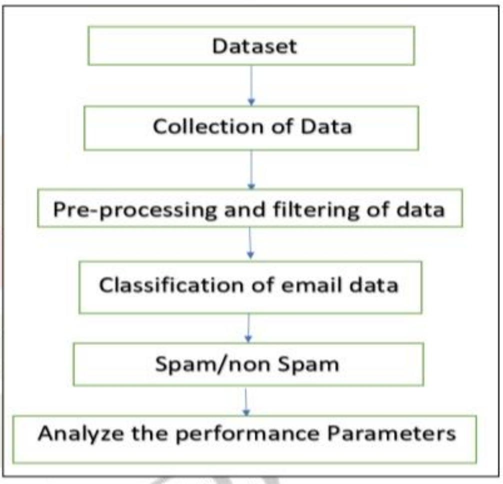

Dataset
Collection of Data
Pre-processing and filtering of data
Classification of email data
Spam/non Spam
Analyze the performance Parameters

Fig2. Flow chart

In our system, according to fig1, incoming E-mails will be pre-processed and features are extracted and the features extracted are given to the training phase. And Email corpus is the text file of the email from which features are extracted and rough set rules are generated. Then using machine learning models like naïve bayes and KNN, emails are classified as spam and non-spam. The accuracy, precision and recall of naïve bayes and KNN will be compared. Fig2 represents flow chart of our system. In which emails are collected from the dataset and it is pre-processed and filtered and the email will be detected as spam or non-spam. And finally, we will conclude that which algorithm is more accurate.

V. CONCLUSION

Email has been the most important medium of communication nowadays; through internet connectivity any message can be delivered to all aver the world. More than 270 billion emails are exchanged daily, about 57% of these are just spam emails. Spam emails, also known as non-self, are undesired commercial or malicious emails, which affects or hacks personal information like bank, related to money or anything that causes destruction to single individual or a corporation or a group of people.

Hence this system is designed in such a way that it detects unsolicited and unwanted emails and prevents them hence helping in reducing the spam message which would be of great benefit to individuals as well as to the company. In the future this system can be implemented by using different algorithms and also more features can be added to the existing system. We will implement email monitoring system using Machine Learning Algorithms such as Naive Bayes Algorithm and KNN (K-Nearest Neighbor) Algorithm and compare which is better.

REFERENCES

[1] "Email based Spam Detection"
Thashina Sultana, K A Sapnaz, Fathima Sana, Mrs. Jamedar Najath, Dept. of Computer Science and Engineering, Yenepoya Institute of Technology
Moodbidri, India

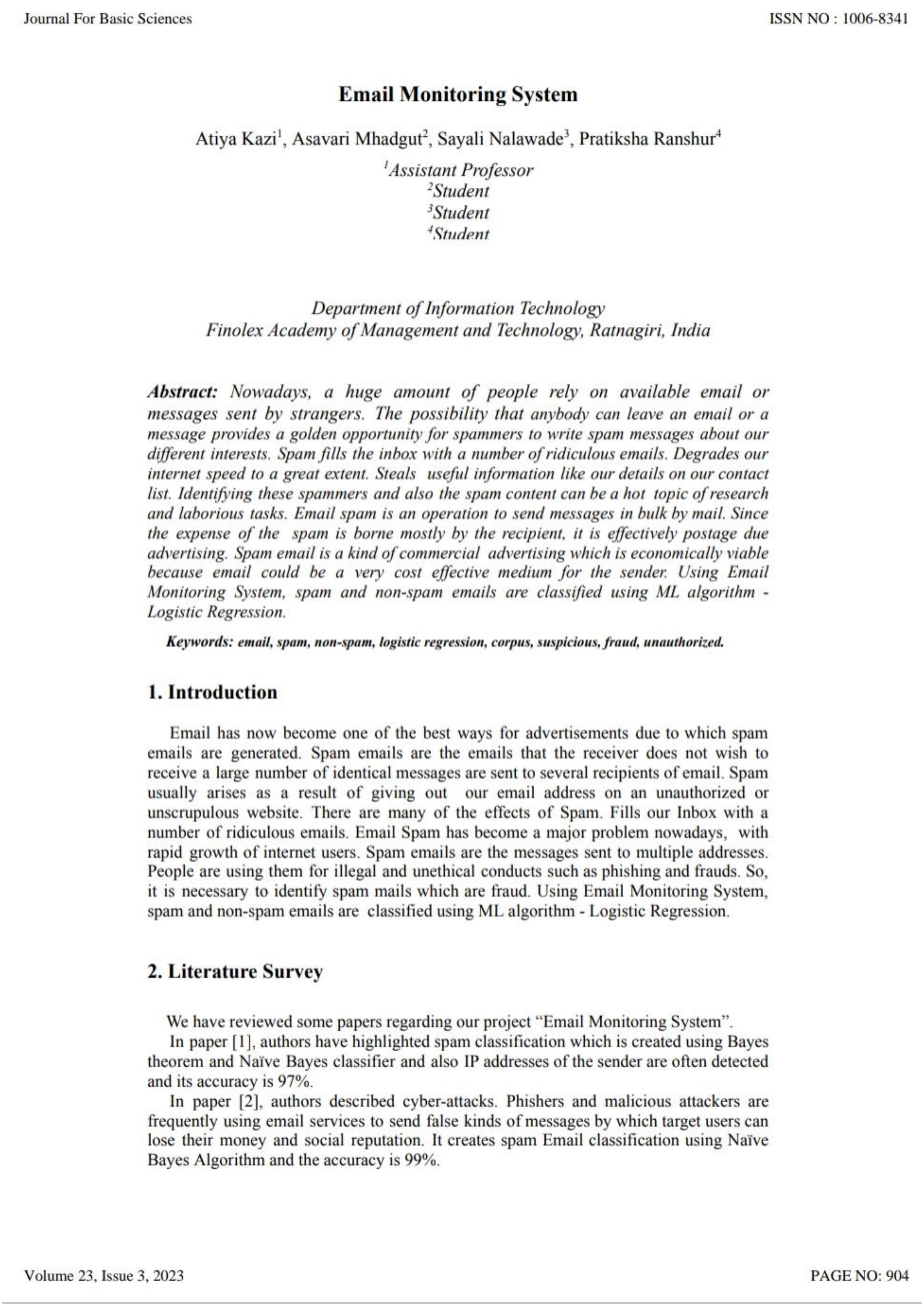
[2] "Email Spam Detection using Naive Bayes Classifier"
Megha Tope, ME Student, Computer Science and Engineering, CSMSS College of Engineering, Aurangabad, India

[3] "A Survey of Existing E-Mail Spam Filtering Methods" Considering Machine Learning Techniques
By Hanif Bhuiyan, Akm Ashiquzzaman, Tamanna Islam Juthi, Suzit Biswas & Jinat Ara, Southeast University

[4] "Email Spam Filtering Using Machine Learning Based Xgboost Classifier Method", 1p.U. Anitha, 2dr.C.V. Guru Rao, 3dr. D. Suresh Babu, 1PhD Scholar, Dept of Computer Science and Engineering, JNTU, Hyderabad
2DIRECTOR, S.R. Engineering College, Warangal, Telangana, India, 3H.O.D, Department of Computer Science and Engineering, Kakatiya Government College.

[5] "Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges", Naeem Ahmed, Rashid Amin, Hamza Aldabbas, Deepika Koundal, Bader Alouffi, and Tariq Shah

8.2 Journal



Email Monitoring System

Atiya Kazi[1], Asavari Mhadgut[2], Sayali Nalawade[3], Pratiksha Ranshur[4]

[1]Assistant Professor
[2]Student
[3]Student
[4]Student

Department of Information Technology
Finolex Academy of Management and Technology, Ratnagiri, India

Abstract: *Nowadays, a huge amount of people rely on available email or messages sent by strangers. The possibility that anybody can leave an email or a message provides a golden opportunity for spammers to write spam messages about our different interests. Spam fills the inbox with a number of ridiculous emails. Degrades our internet speed to a great extent. Steals useful information like our details on our contact list. Identifying these spammers and also the spam content can be a hot topic of research and laborious tasks. Email spam is an operation to send messages in bulk by mail. Since the expense of the spam is borne mostly by the recipient, it is effectively postage due advertising. Spam email is a kind of commercial advertising which is economically viable because email could be a very cost effective medium for the sender. Using Email Monitoring System, spam and non-spam emails are classified using ML algorithm - Logistic Regression.*

Keywords: ***email, spam, non-spam, logistic regression, corpus, suspicious, fraud, unauthorized.***

1. Introduction

Email has now become one of the best ways for advertisements due to which spam emails are generated. Spam emails are the emails that the receiver does not wish to receive a large number of identical messages are sent to several recipients of email. Spam usually arises as a result of giving out our email address on an unauthorized or unscrupulous website. There are many of the effects of Spam. Fills our Inbox with a number of ridiculous emails. Email Spam has become a major problem nowadays, with rapid growth of internet users. Spam emails are the messages sent to multiple addresses. People are using them for illegal and unethical conducts such as phishing and frauds. So, it is necessary to identify spam mails which are fraud. Using Email Monitoring System, spam and non-spam emails are classified using ML algorithm - Logistic Regression.

2. Literature Survey

We have reviewed some papers regarding our project "Email Monitoring System".

In paper [1], authors have highlighted spam classification which is created using Bayes theorem and Naïve Bayes classifier and also IP addresses of the sender are often detected and its accuracy is 97%.

In paper [2], authors described cyber-attacks. Phishers and malicious attackers are frequently using email services to send false kinds of messages by which target users can lose their money and social reputation. It creates spam Email classification using Naïve Bayes Algorithm and the accuracy is 99%.

In paper [3], the system Uses Machine Learning techniques like Naïve Bayes, KNN, SVM and summarizes the overall scenario regarding Accuracy Rate.

In paper [4], It compares the implemented algorithm of XGboosting Classifier with the existing classifier of SVM, Naïve Bayes and its accuracy is 95% .

In paper [5], It implements spam detection by using Machine learning and deep Learning Techniques like Naïve Bayes, Decision Tree, Neural Networks and random forest.

3. Existing System

The Existing Email system has a weak spam detection Mechanism. This results in segregation of important emails into spam emails. This can lead to miscommunication or delay of messages. Sometimes, some unimportant or phishing emails are not segregated into spam folders which leads to unnecessary attention to unimportant emails.

Drawbacks of existing system are as follows:

It is the source of viruses. It is capable of harming one's computer and reading out a user's email address book and sending them to a number of people around the world. It can be the source of various spams. These spam mails can fill up the inbox and deletion of these mail consumes a lot of time. It is an informal method of communication. The documents that require signatures are not managed by email.

4. Proposed System

In our system to solve the problem of spam, the email monitoring system is implemented using a machine learning algorithm - Logistic Regression. Using our system, the problem of the existing system will be solved.

There are some advantages of our system:

Ensemble methods on the other hand proved to be useful as they used multiple classifiers for class prediction. Nowadays, lots of emails are sent and received and it is difficult as our project is only able to test emails using a limited amount of corpus. Our project, thus Email monitoring system, is proficient in filtering mails giving to the content of the email and not according to the domain names or any other criteria. Also, our system has good efficiency as well as greater accuracy.

5. Flow Diagram

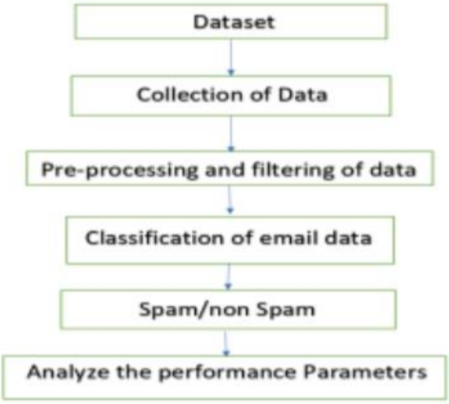

Dataset
Collection of Data
Pre-processing and filtering of data
Classification of email data
Spam/non Spam
Analyze the performance Parameters

Figure 1. Flow Diagram

Figure 1 represents the flow chart of our system. In which emails are collected from the dataset and it is pre-processed and filtered and the email will be detected as spam or non-spam.

6. Methodology

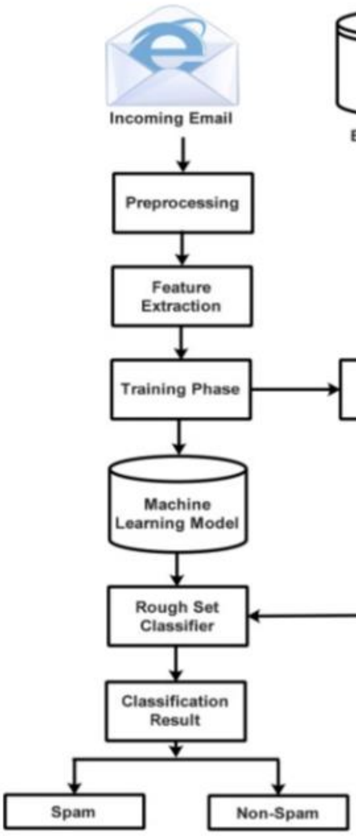
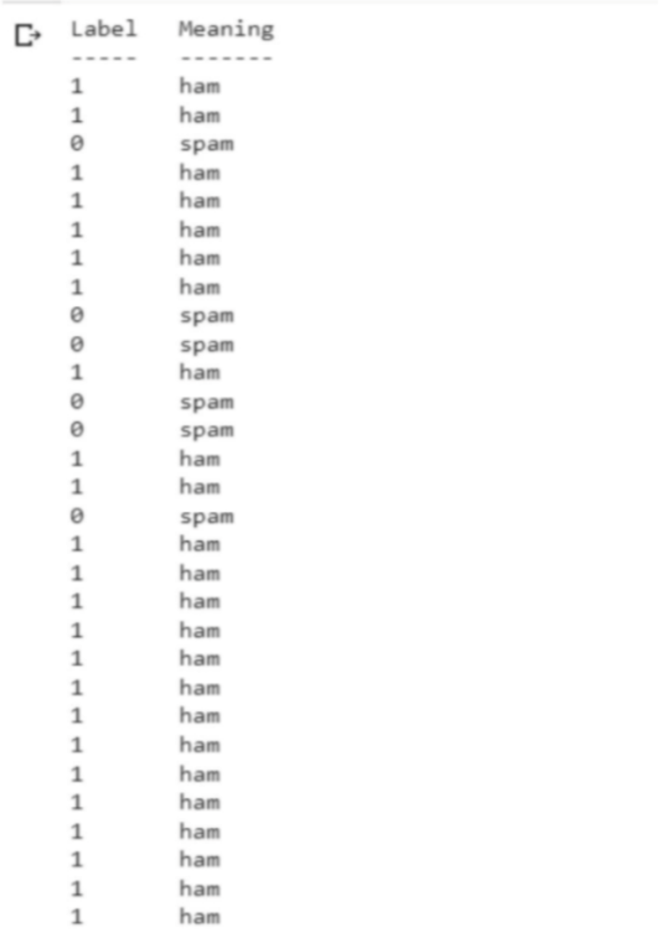


Figure 2. System Architecture

In our system, according to Figure 2, incoming Emails will be pre-processed and features are extracted and the features extracted are given to the training phase. And Email corpus is the text file of the email from which features are extracted and rough set rules are generated. Then using the machine learning model - Logistic Regression, emails are classified as spam and non-spam.

7. Result



```
Label   Meaning
-----   -------
1       ham
1       ham
0       spam
1       ham
1       ham
1       ham
1       ham
1       ham
0       spam
0       spam
1       ham
0       spam
0       spam
1       ham
1       ham
0       spam
1       ham
1       ham
1       ham
1       ham
1       ham
1       ham
1       ham
1       ham
1       ham
1       ham
1       ham
1       ham
1       ham
1       ham
```

8. Conclusion

Email has been the most important medium of communication nowadays; through internet connectivity any message can be delivered to all over the world. More than 270 billion emails are exchanged daily, about 57% of these are just spam emails. Spam emails, also known as non-self, are undesired commercial or malicious emails, which affects or hacks personal information like bank, related to money or anything that causes destruction to a single individual or a corporation or a group of people.

Hence this system is designed in such a way that it detects unsolicited and unwanted emails and prevents them hence helping in reducing the spam message which would be of great benefit to individuals as well as to the company. In the future this system can be implemented by using different algorithms and also more features can be added to the existing system. We will implement an email monitoring system using Machine Learning Algorithm - Logistic Regression.

9. References

[1] “Email based Spam Detection”
Thashina Sultana, K A Sapnaz, Fathima Sana, Mrs. Jamedar Najath, Dept. of Computer Science and Engineering, Yenepoya Institute of Technology Moodbidri, India.

[2] “Email Spam Detection using Naive Bayes Classifier”
Megha Tope, ME Student, Computer Science and Engineering, CSMSS College of Engineering, Aurangabad, India.

[3] “A Survey of Existing E-Mail Spam Filtering Methods”
Considering Machine Learning Techniques By Hanif Bhuiyan, Akm Ashiquzzaman, Tamanna Islam Juthi, Suzit Biswas & Jinat Ara, Southeast University.

[4] “Email Spam Filtering Using Machine Learning Based Xgboost Classifier Method”,
1p.U. Anitha, 2dr.C.V. Guru Rao, 3dr. D. Suresh Babu, 1PhD Scholar, Dept of Computer Science and Engineering, JNTU, Hyderabad 2DIRECTOR, S.R. Engineering College, Warangal, Telangana, India, 3H.O.D, Department of Computer Science and Engineering, Kakatiya Government College.

[5] “Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges”,
Naeem Ahmed, Rashid Amin, Hamza Aldabbas, Deepika Koundal, Bader Alouffi, and Tariq Shah.

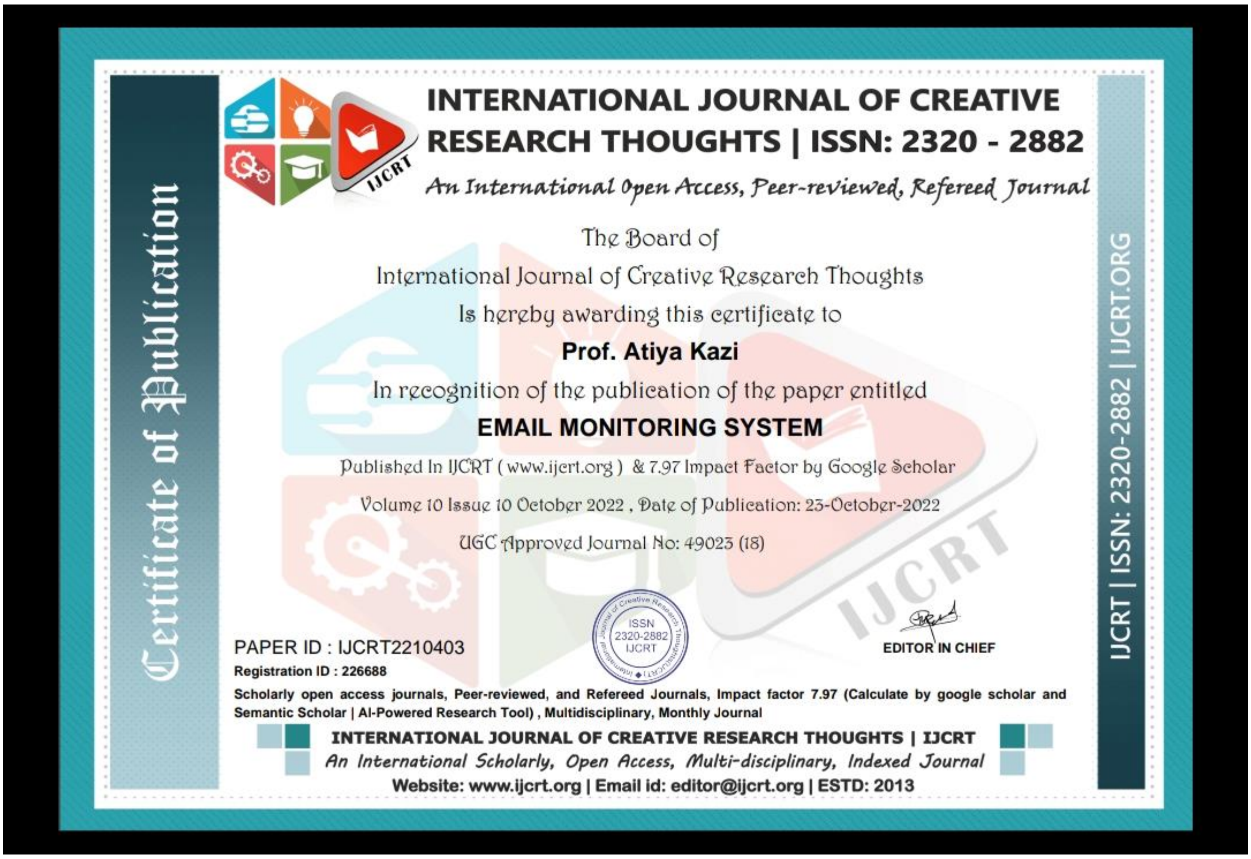
[6] “Identification of Spam Email Based on Information from Email Header”,
Shukor Bin Abd Razak, Ahmad Fahrulrazie Bin Mohamad 13th International Conference on Intelligent Systems Design and Applications (ISDA), 2013.

[7] “E-Mail Spam Detection Based on Part of Speech Tagging”,
Mohammed Reza Parsei, Mohammed Salehi 2nd International Conference on Knowledge Based Engineering and Innovation (KBEI), 2015.

[8] "Email classification research trends: Review and open issues",
Mujtaba, Ghulam, et al. IEEE Access 5 (2017).

[9] "Spam/ham e-mail classification using machine learning methods based on bag of words technique",
Sahın, Esra, Murat Aydos, and Fatih Orhan, 2018 26th Signal Processing and Communications Applications Conference (SIU). IEEE, 2018.

8.3 Certificates



Certificate of Publication

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN: 2320 - 2882

An International Open Access, Peer-reviewed, Refereed Journal

The Board of

International Journal of Creative Research Thoughts

Is hereby awarding this certificate to

Prof. Atiya Kazi

In recognition of the publication of the paper entitled

EMAIL MONITORING SYSTEM

Published In IJCRT (www.ijcrt.org) & 7.97 Impact Factor by Google Scholar

Volume 10 Issue 10 October 2022 , Date of Publication: 23-October-2022

UGC Approved Journal No: 49023 (18)

ISSN 2320-2882 IJCRT

EDITOR IN CHIEF

PAPER ID : IJCRT2210403

Registration ID : 226688

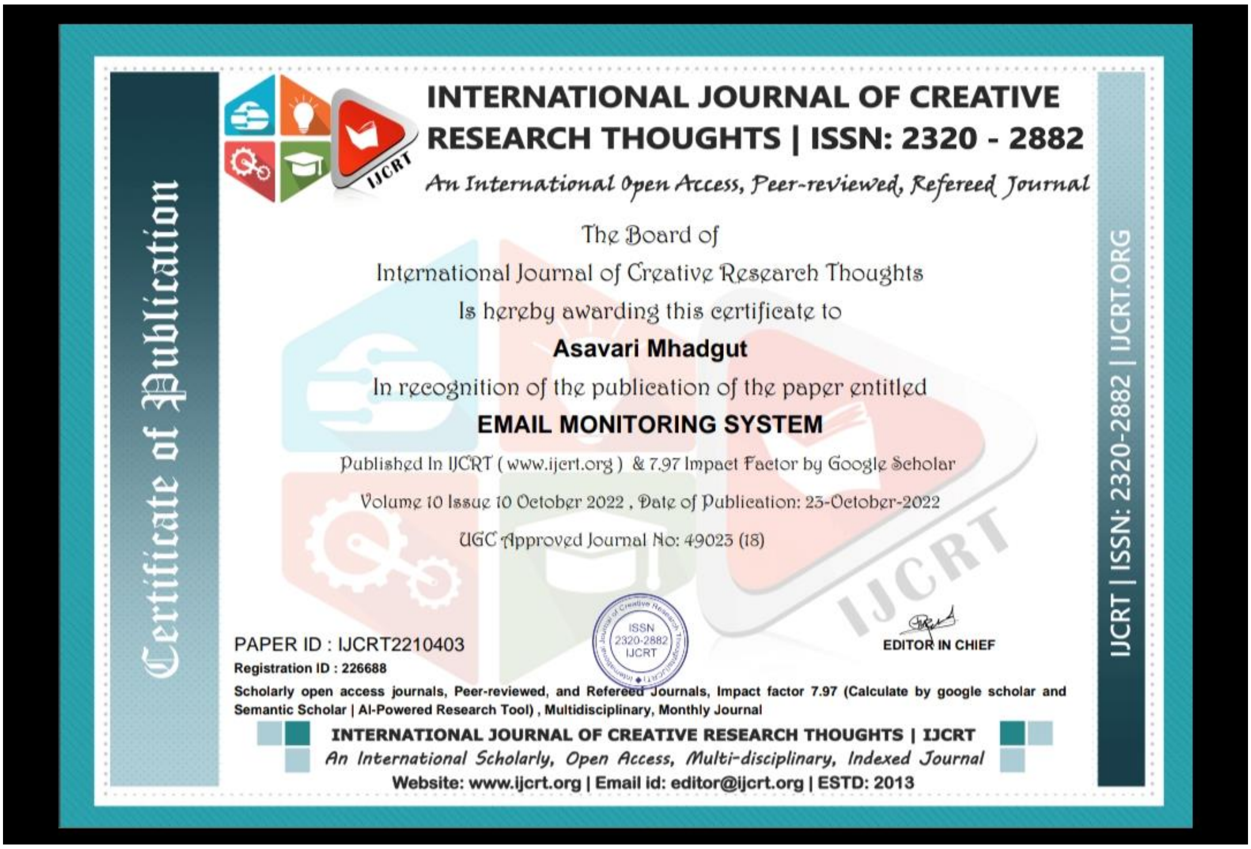
Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | IJCRT

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijcrt.org | Email id: editor@ijcrt.org | ESTD: 2013

IJCRT | ISSN: 2320-2882 | IJCRT.ORG



Certificate of Publication

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN: 2320 - 2882

An International Open Access, Peer-reviewed, Refereed Journal

The Board of

International Journal of Creative Research Thoughts

Is hereby awarding this certificate to

Asavari Mhadgut

In recognition of the publication of the paper entitled

EMAIL MONITORING SYSTEM

Published In IJCRT (www.ijcrt.org) & 7.97 Impact Factor by Google Scholar

Volume 10 Issue 10 October 2022 , Date of Publication: 23-October-2022

UGC Approved Journal No: 49023 (18)

ISSN 2320-2882 IJCRT

EDITOR IN CHIEF

PAPER ID : IJCRT2210403

Registration ID : 226688

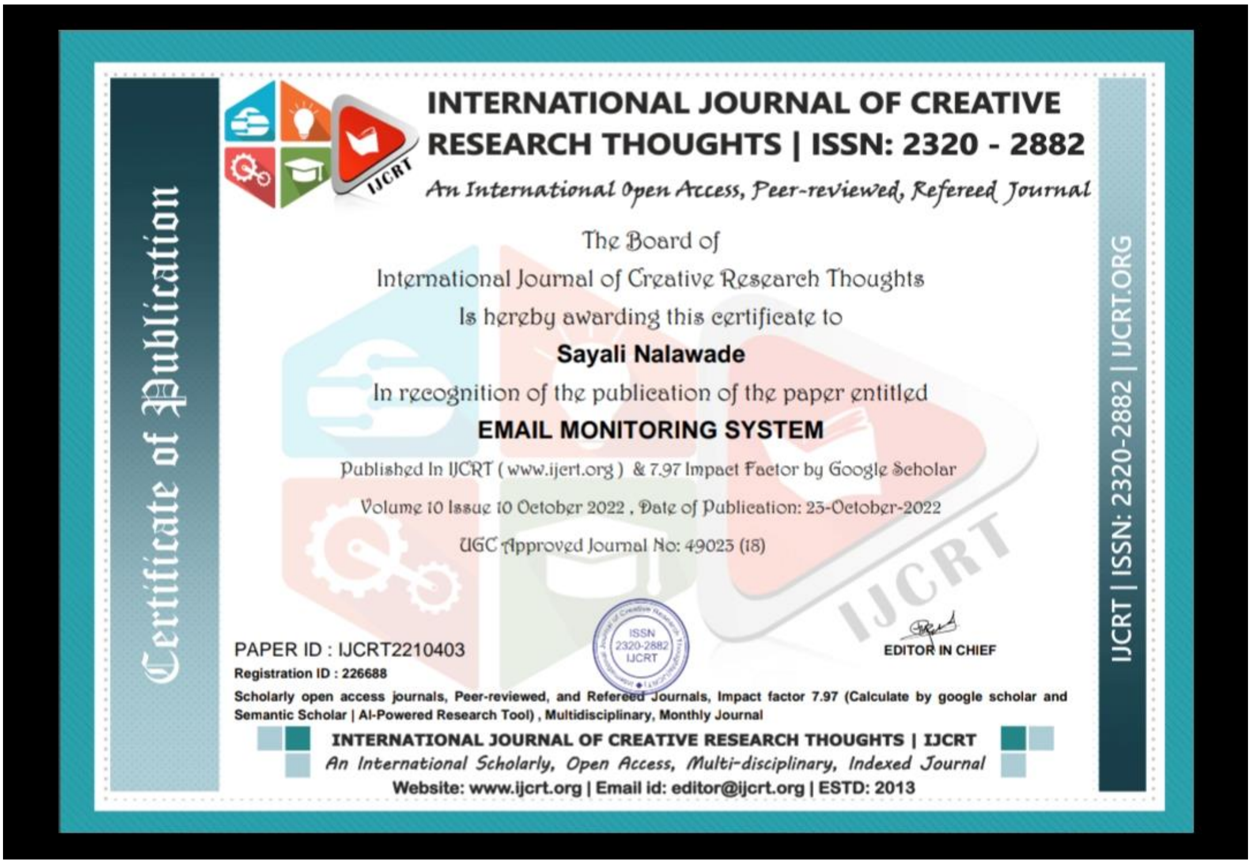
Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | IJCRT

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijcrt.org | Email id: editor@ijcrt.org | ESTD: 2013

IJCRT | ISSN: 2320-2882 | IJCRT.ORG



Certificate of Publication

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN: 2320 - 2882

An International Open Access, Peer-reviewed, Refereed Journal

The Board of
International Journal of Creative Research Thoughts
Is hereby awarding this certificate to

Sayali Nalawade

In recognition of the publication of the paper entitled

EMAIL MONITORING SYSTEM

Published In IJCRT (www.ijcrt.org) & 7.97 Impact Factor by Google Scholar

Volume 10 Issue 10 October 2022 , Date of Publication: 23-October-2022

UGC Approved Journal No: 49023 (18)

PAPER ID : IJCRT2210403

Registration ID : 226688

ISSN 2320-2882 IJCRT

EDITOR IN CHIEF

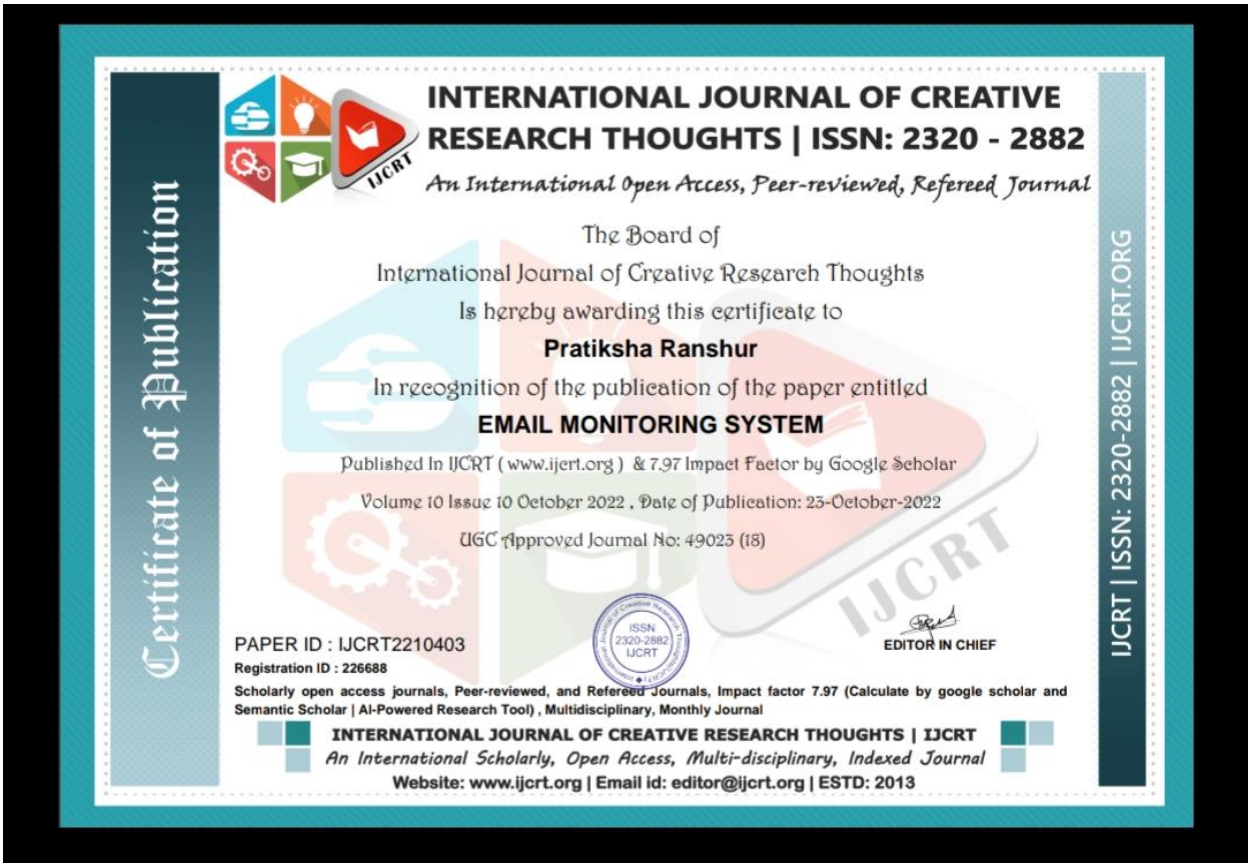
Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | IJCRT

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijcrt.org | Email id: editor@ijcrt.org | ESTD: 2013

IJCRT | ISSN: 2320-2882 | IJCRT.ORG



Certificate of Publication

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN: 2320 - 2882

An International Open Access, Peer-reviewed, Refereed Journal

The Board of
International Journal of Creative Research Thoughts
Is hereby awarding this certificate to

Pratiksha Ranshur

In recognition of the publication of the paper entitled

EMAIL MONITORING SYSTEM

Published In IJCRT (www.ijcrt.org) & 7.97 Impact Factor by Google Scholar

Volume 10 Issue 10 October 2022 , Date of Publication: 23-October-2022

UGC Approved Journal No: 49023 (18)

PAPER ID : IJCRT2210403

Registration ID : 226688

ISSN 2320-2882 IJCRT

EDITOR IN CHIEF

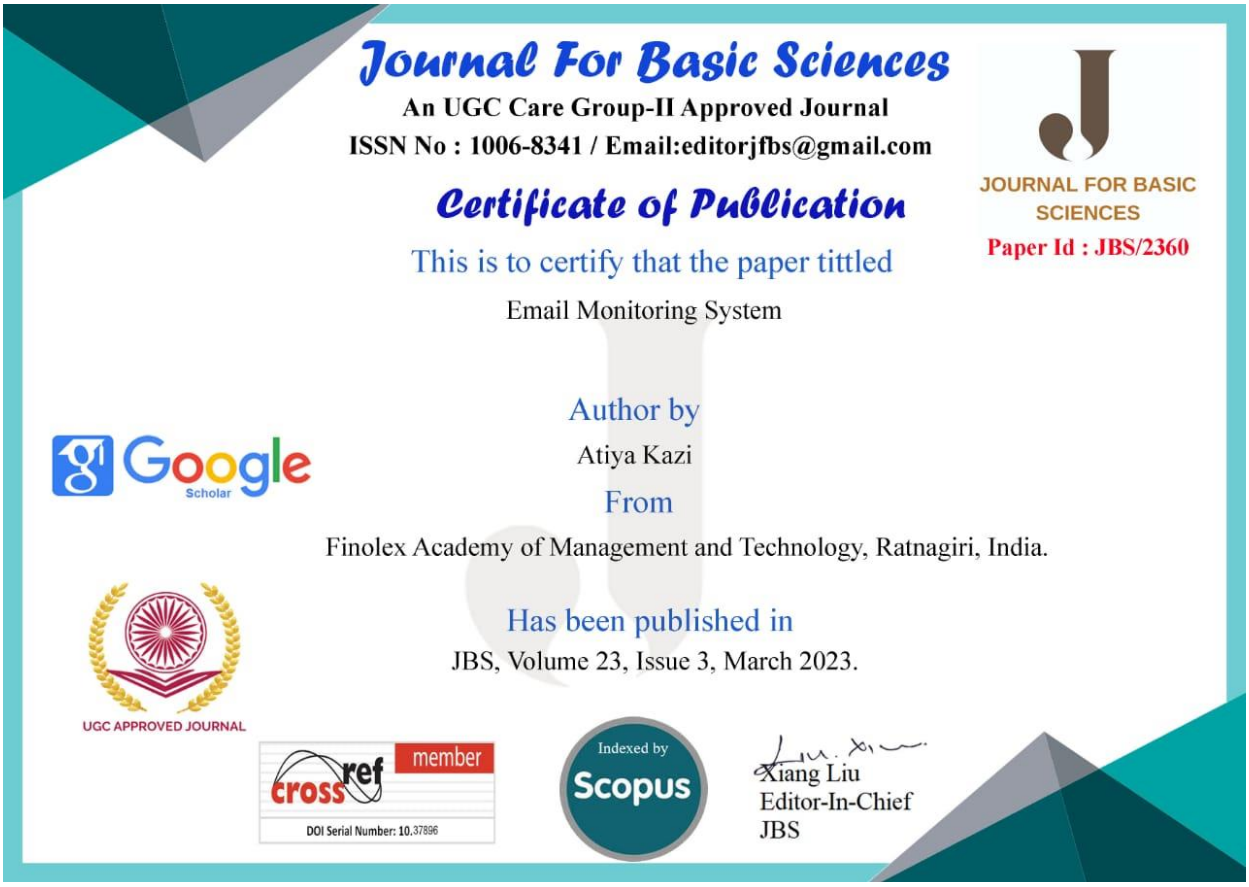
Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | IJCRT

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijcrt.org | Email id: editor@ijcrt.org | ESTD: 2013

IJCRT | ISSN: 2320-2882 | IJCRT.ORG



Journal For Basic Sciences

An UGC Care Group-II Approved Journal

ISSN No : 1006-8341 / Email:editorjfbs@gmail.com

JOURNAL FOR BASIC SCIENCES

Paper Id : JBS/2360

Certificate of Publication

This is to certify that the paper tittled

Email Monitoring System

Author by

Atiya Kazi

From

Finolex Academy of Management and Technology, Ratnagiri, India.

Has been published in

JBS, Volume 23, Issue 3, March 2023.

Google Scholar

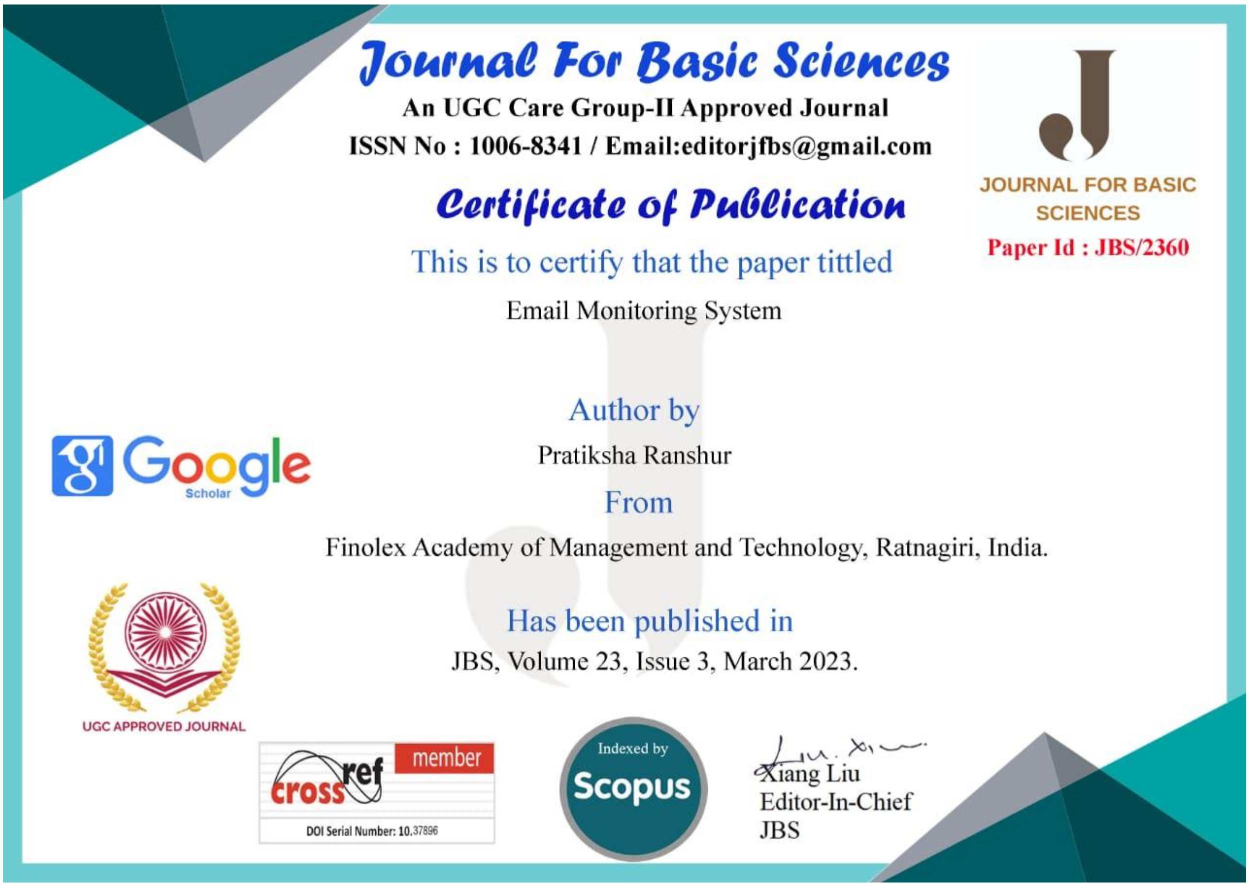
UGC APPROVED JOURNAL

crossref member

DOI Serial Number: 10.37896

Indexed by Scopus

Xiang Liu
Editor-In-Chief
JBS



Journal For Basic Sciences

An UGC Care Group-II Approved Journal

ISSN No : 1006-8341 / Email:editorjfbs@gmail.com

JOURNAL FOR BASIC SCIENCES

Paper Id : JBS/2360

Certificate of Publication

This is to certify that the paper tittled

Email Monitoring System

Author by

Pratiksha Ranshur

From

Finolex Academy of Management and Technology, Ratnagiri, India.

Has been published in

JBS, Volume 23, Issue 3, March 2023.

Google Scholar

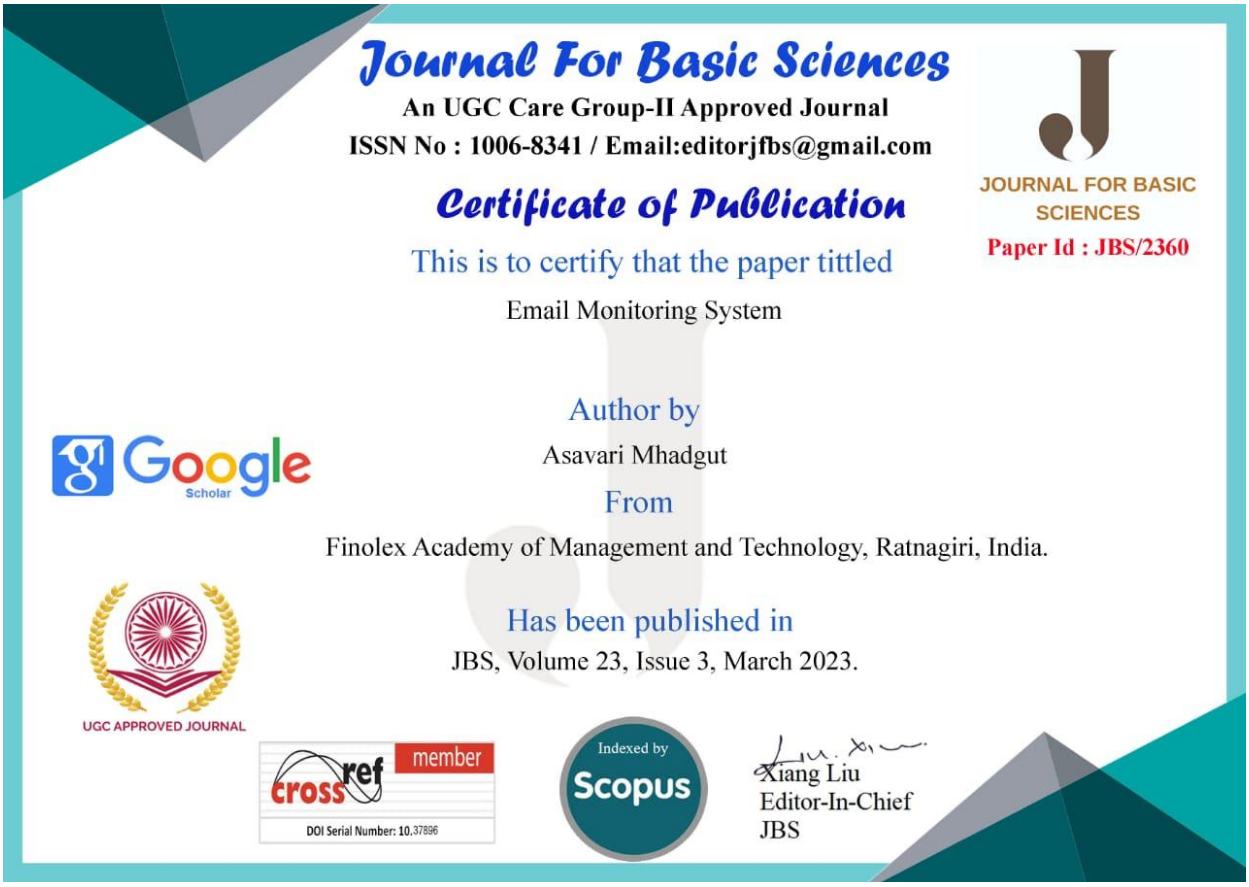
UGC APPROVED JOURNAL

crossref member

DOI Serial Number: 10.37896

Indexed by Scopus

Xiang Liu
Editor-In-Chief
JBS



Journal For Basic Sciences

An UGC Care Group-II Approved Journal

ISSN No : 1006-8341 / Email:editorjfbs@gmail.com

JOURNAL FOR BASIC SCIENCES

Paper Id : JBS/2360

Certificate of Publication

This is to certify that the paper tittled

Email Monitoring System

Author by

Asavari Mhadgut

From

Finolex Academy of Management and Technology, Ratnagiri, India.

Has been published in

JBS, Volume 23, Issue 3, March 2023.

Google Scholar

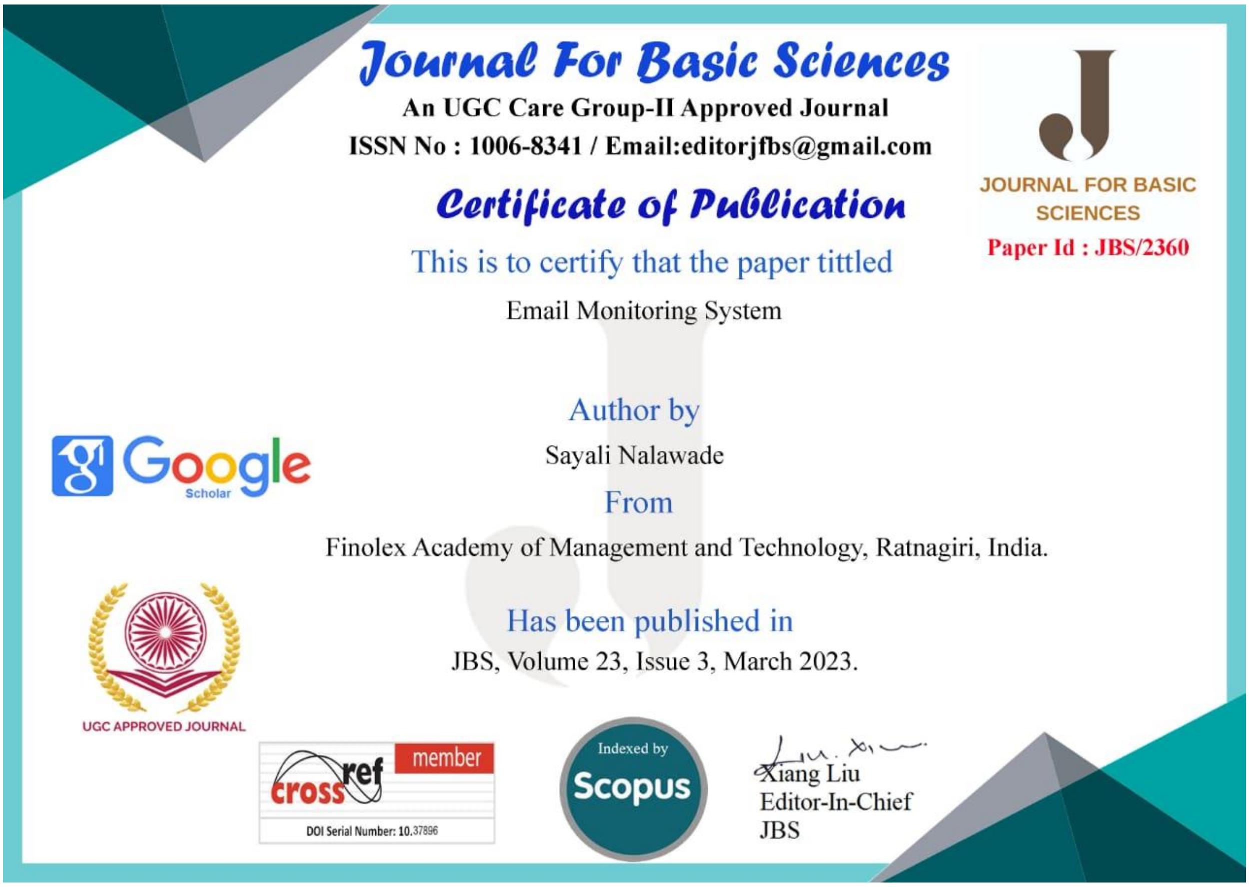
UGC APPROVED JOURNAL

crossref member

DOI Serial Number: 10.37896

Indexed by Scopus

Xiang Liu
Editor-In-Chief
JBS



Journal For Basic Sciences

An UGC Care Group-II Approved Journal

ISSN No : 1006-8341 / Email:editorjfbs@gmail.com

JOURNAL FOR BASIC SCIENCES

Paper Id : JBS/2360

Certificate of Publication

This is to certify that the paper tittled

Email Monitoring System

Author by

Sayali Nalawade

From

Finolex Academy of Management and Technology, Ratnagiri, India.

Has been published in

JBS, Volume 23, Issue 3, March 2023.

Google Scholar

UGC APPROVED JOURNAL

crossref member

DOI Serial Number: 10.37896

Indexed by Scopus

Xiang Liu
Editor-In-Chief
JBS

Chapter 9

Appendix

Acknowledgement

We are using this opportunity to express our gratitude to everyone who supported us throughout the course of this project. We are thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. We are sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

We owe our deep gratitude to our project guide Prof Atiya Kazi, who took keen interest in our project work and guided us all along, till the completion of our project work by providing all the necessary information. We are thankful to and fortunate enough to get constant encouragement, support and guidance from teaching staff of Department of Information Technology which helped us in successfully completing our project work