IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

Detecting Spam Email with Machine Learning Optimized with Bio-Inspired Metaheuristic Algorithms

Vatan Koshti[1], Aditi Gaherwar[2], Twinkle Ramteke[3], Yogeshwari Durgam[4], Prof. Madhavi Sadu[5]
Students, Department of Information Technology[1,2,3,4]
Professor, Department of Information Technology[5]
Rajiv Gandhi College of Engineering Research and Technology, Chandrapur, India

Abstract: Electronic mail has eased communication methods for many organizations as well as individuals. Spammers use this strategy to make fraudulent gains by sending unsolicited emails. This project aims to present a method for detection of spam emails with machine learning algorithms that are optimized with bio-inspired methods. A literature review is carried to explore the efficient methods applied on different datasets to achieve good results. The bio-inspired algorithms like Particle Swarm Optimization and Genetic Algorithm were implemented to optimize the performance of classifiers. Multinomial Naïve Bayes with Genetic Algorithm performed the best overall. The comparison of our results with other machine learning and bio-inspired models to show the best suitable model is also discussed.

Keywords: Bio inspired algorithm, Particle Swarm Optimization algorithm

I. INTRODUCTION

Nowadays, Emails are used in almost every field, from business to 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 used to harm any user by wasting his/her time, computing resources, and stealing valuable information. The ratio of spam emails is increasing rapidly day by day.

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. Creating a fake profile and email account is much easy for the spammers, they pretend like a genuine person in their spam emails, these spammers target those peoples who are not aware about these frauds. Machine learning models have been utilized for multiple purposes in the field of computer science from resolving a network traffic issue to detecting a malware.Spam e-mail are message randomly sent to multiple addresses by all sorts of groups, but mostly lazy advertisers and criminals who wish to lead you to phishing sites.Recently unsolicited commercial/bulk e-mail also known as spam, become a big trouble over the internet. Spam is waste of time, storage space and communication bandwidth.So, it is needed to Identify those spam mails which are fraud, this project will identify those spam by using techniques of machine learning Optimized with Bio-Inspired Metaheuristic Algorithms", this project will discuss the machine learning algorithms and apply all these algorithms on our data sets and best algorithm is selected for the email spam detection having best precision and accuracy.

The Scope of this project aims to achieve the following:

1) To explore machine learning algorithms for the spam detection problem.
2) To investigate the workings of the algorithms with the acquired datasets.
3) To implement the bio-inspired algorithms.
4) To test and compare the accuracy of base models with bio-inspired implementation.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

II. LITERATURE SURVEY

1) A support vector machine based Naive Bayes algorithm for spam filtering

AUTHORS: W. Feng, J. Sun, L. Zhang, C. Cao, and Q. Yang

Naive Bayes classifiers are widely used to filter spam emails, however, the strong independence assumptions between features limit their performance in accurately identifying spams. To address this issue, we proposed a support machine vector based naive Bayes - SVM-NB - filtering system. The SVM-NB first constructs an optimal separating hyperplane that divides samples in the training set into two categories. For samples located nearby the hyperplane, if they are in different categories, one of them will be eliminated from the training set. In this way, the dependence between samples is reduced and the entire training sample space is simplified. With the trimmed training set, the naive Bayes algorithm is applied to classify emails in the test set. The SVM-NB system is evaluated with the dataset obtained from DATAMALL. Experiment results demonstrate that SVM-NB can achieve a higher spam-detection accuracy and a faster classification speed.

2) Machine learning for email spam filtering: Review, approaches and open research problems

AUTHORS: E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa

The upsurge in the volume of unwanted emails called spam has created an intense need for the development of more dependable and robust antispam filters. Machine learning methods of recent are being used to successfully detect and filter spam emails. We present a systematic review of some of the popular machine learning based email spam filtering approaches. Our review covers survey of the important concepts, attempts, efficiency, and the research trend in spam filtering. The preliminary discussion in the study background examines the applications of machine learning techniques to the email spam filtering process of the leading internet service providers (ISPs) like Gmail, Yahoo and Outlook emails spam filters. Discussion on general email spam filtering process, and the various efforts by different researchers in combating spam through the use machine learning techniques was done. Our review compares the strengths and drawbacks of existing machine learning approaches and the open research problems in spam filtering. We recommended deep leaning and deep adversarial learning as the future techniques that can effectively handle the menace of spam emails.

3) Machine learning methods for spam E-Mail classification

AUTHORS: W. Awad and S. ELseuo

The increasing volume of unsolicited bulk e-mail (also known as spam) has generated a need for reliable anti-spam filters. Machine learning techniques now days used to automatically filter the spam e-mail in a very successful rate. In this paper we review some of the most popular machine learning methods (Bayesian classification, k-NN, ANNs, SVMs, Artificial immune system and Rough sets) and of their applicability to the problem of spam Email classification. Descriptions of the algorithms are presented, and the comparison of their performance on the SpamAssassin spam corpus is presented.

4) Classifying unsolicited bulk email (UBE) using Python machine learning techniques

AUTHORS: S. Mohammed, O. Mohammed, and J. Fiaidhi

Email has become one of the fastest and most economical forms of communication. However, the increase of email users has resulted in the dramatic increase of spam emails during the past few years. As spammers always try to find a way to evade existing filters, new filters need to be developed to catch spam. Generally, the main tool for email filtering is based on text classification. A classifier then is a system that classifies incoming messages as spam or legitimate (ham) using classification methods. The most important methods of classification utilize machine learning techniques. There are a plethora of options when it comes to deciding how to add a machine learning component to a python email classification. This article describes an approach for spam filtering using Python where the interesting spam or ham words (spam-ham lexicon) are filtered first from the training dataset and then this lexicon is used to generate the training and testing tables that are used by variety of data mining algorithms. Our experimentation using one dataset reveals the affectivity of the Naïve Bayes and the SVM classifiers for spam filtering.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

5) Hybrid decision tree and logistic regression classifier for email spam detection

AUTHORS: A. Wijaya and A. Bisri

Email spam is an increasing problem because it disrupting and time consuming for user, since the easy and cheap of sending email. Email Spam filtering can be done with a binary classification with machine learning as classifier. To date, email spam detection still challenging since the email spam still happens a lot and the detection still need improvement. Decision Tree (DT) is one of famous classifier since DT able to handle nominal and numerical attributes and increasing the efficiency of computing. However, DT has a weakness in over-sensitivity to the training set and the noise data or instance that can degrade the performance. In this study, we propose hybrid combination Logistic Regression (LR) and DT for email spam detection. LR is used for reduce noisy data or instance before data feed to DT induction. Noisy data reducing is done by LR by filtering correct prediction with certain false negative threshold. In this study, Spambase dataset is used to evaluate the proposed method. From the experiment, the result shows that proposed method yield impressive and promising result with the accuracy is 91.67%. It can be concluded that LR able to improve DT performance by reducing noisy data.

Spam

Unsolicited usually commercial messages (such as emails, text messages, or Internet postings) sent to a large number of recipients or posted in a large number of places.

What is spam how it is harmful

Most spam is irritating and time-consuming, but some spam is positively dangerous to handle. Usually email scams are trying to get you to give up your bank details so that the fraudsters can either withdraw money or steal your identity. Such messages include phishing scams and advanced fee fraud.

III. INPUT DESIGN AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

What data should be given as input?

How the data should be arranged or coded?

The dialog to guide the operating personnel in providing input.

Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoiderrors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2.It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easilyand effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
2. Select methods for presenting information.
3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

IV. SYSTEM ANALYSIS:

EXISTING SYSTEM:

- S. L. Marie-Sainte and N. Alalyani used the Firefly algorithm with SVM. The researchers experimented with the Arabic text with feature selection. The paper concluded that the proposed method outperforms the SVM itself.
- E. A. Natarajan, S. Subramanian, and K. Premalatha proposed Enhanced Cuckoo Search (ECS) for bloom filter optimization. This is where the weight of the spam word is considered. It was concluded that their proposed optimization technique of ECS outperforms the normal Cuckoo search.

DISADVANTAGES OF EXISTING SYSTEM:

- Spam detection rules are set up and are in constant need of manual updating thus consuming time and resources.
- The problem of selecting the set of attributes is NP-hard.
- Less Accuracy.
- More time taking process.

V. PROPOSED SYSTEM

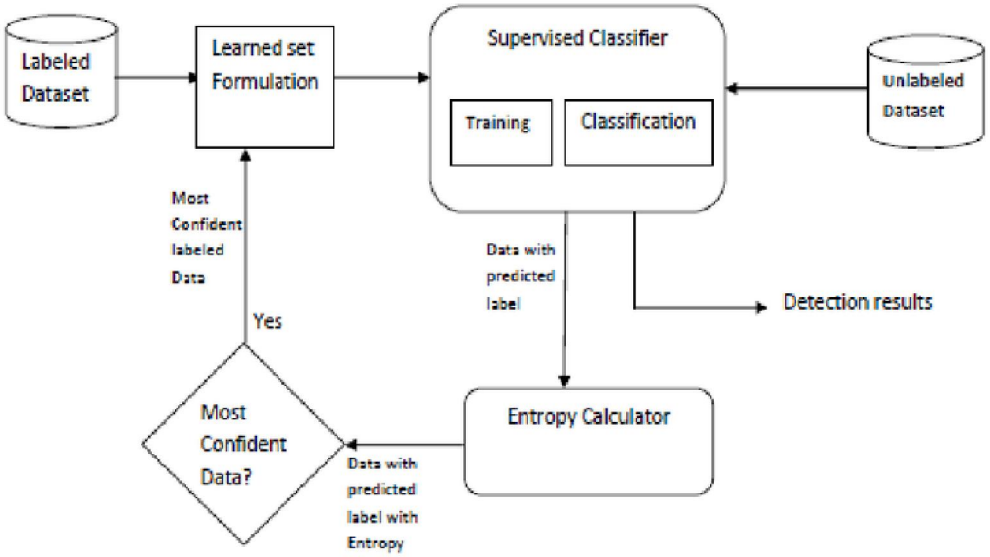
- The proposed spam detection to resolve the issue of the spam classification problem can be further experimented by feature selection or automated parameter selection for the models.
- The paper aims to achieve the following objectives
- To explore machine learning algorithms for the spam detection problem.
- To investigate the workings of the algorithms with the acquired datasets.
- To test and compare the accuracy of base models with bio-inspired implementation.
- To implement the framework using Python.
- Scikit-Learn library will be explored to perform the experiments with Python, and this will enable to edit the models, conduct pre-processing and calculate the results. The program scripts will be implemented further with the optimization techniques and compared with the base results i.e with default parameters.
- The spam detection engine should be able to take email datasets as input and with the help of text mining and optimized supervised algorithms, it should be able to classify the email as ham or spam

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

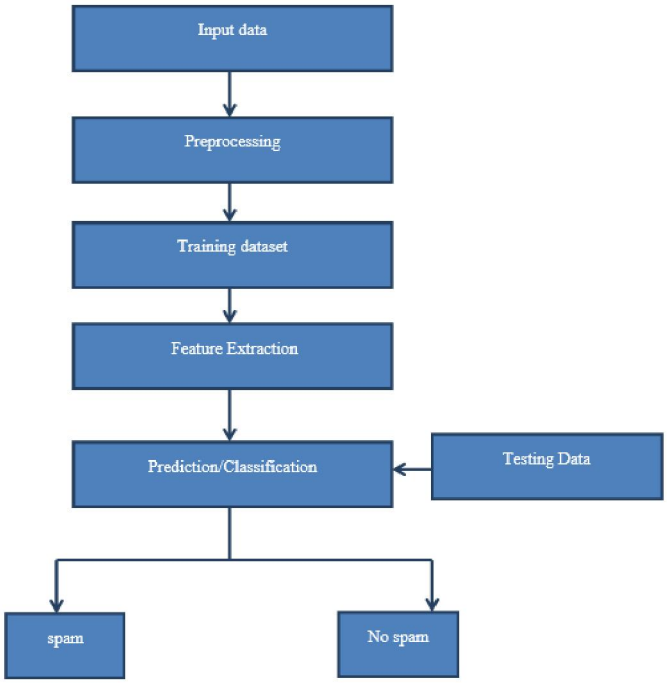
ADVANTAGES OF PROPOSED SYSTEM:

- Machine learning makes it easier because it learns to recognise the unsolicited emails (spam) and legitimate emails (ham) automatically and then applies those learned instructions to unknown incoming emails.
- Achieved good results overall for the spam datasets.
- Better Performance with better speed compared to the existing systems.

VI. SYSTEM DESIGN



DATA FLOW DIAGRAM:



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

VII. SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. Theexpenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

VIII. IMPLEMENTATION

MODULES:

- Data Collection
- Dataset
- Data Preparation
- Model Selection
- Analyse and Prediction
- Accuracy on test set
- Saving the Trained Model

MODULES DESCSRIPTION:

Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.There are several techniques to collect the data, like web scraping, manual interventions and etc.The dataset used in this spam-mails Detectiontaken fromKaggle

Link: https://www.kaggle.com/venky73/spam-mails-dataset

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

Dataset:

The dataset consists of 2913 individual data. There are 3 columns in the dataset, which are described below

Id: unique id

Labels : Labels of Emails which can be either Spam or no spam

1: spam

0: no spam

Text: Emails data

Data Preparation:

We will transform the data. by getting rid of missing data and removing some columns. First, we will create a list of column names that we want to keep or retain.

Next, we drop or remove all columns except for the columns that we want to retain.

Finally, we drop or remove the rows that have missing values from the data set.

Steps to follow:

- Removing extra symbols
- Removing punctuations
- Removing the Stop words
- Stemming
- Tokenization
- Feature extractions
- TF-IDF vectorizer
- Counter vectorizer with TF-IDF transformer

Model Selection:

We used Naïve BAYES algorithms

Naïve BAYES algorithms:

Naïve Bayes model is used to resolve classification problems by using probability techniques. The Naïve Bayes algorithm for this article can be denoted as equation

$$P(Class|WORD)=(P(WORD|Class) \times P(Class))/(P(WORD))$$

where WORD is (word1,word2, . . .wordn) from within an uploaded email and 'Class' is either 'Spam' or 'Ham'. The algorithm calculates the probability of a class from the bag of words provided by the program. Where P(Class | WORD) is a posterior probability, P(WORD | Class) is likelihood and P(Class) is the prior probability .

If 'Class' = Spam, the equation could be rewritten to find the spam email from the given words, and this can be further simplified as equation

$$P(Class|WORD) = 0\ n\ i=1\ P(word\_i|Spam) \times P(Spam)/\ P(word\_1, word\_2, . . . word\_n)$$

There are three types of Naïve Bayes algorithms: Multinomial, Gaussian and Bernoulli. Multinomial Naïve Bayes algorithm has been selected to perform the spam email identification because it is text related and outperforms Gaussian and Bernoulli.

Multinomial Naïve Bayes (MNB) classifier uses Multinomial Distribution for each given feature, focusing on term frequency. The Multinomial Naïve Bayes can be denoted as equation.

$$P(p|n) \propto P(p)\ Y\ 1 \leq k \leq nd\ P(tk\ |p)$$

where the number of tokens is represented by nd, n is the number of emails and P(tk |p) is calculated by:

$$P(tk|p) = (count(tk|p) + 1)/\ (count(tp) + |V|)$$

In the equations, P(tk |p) is identified as the conditional probability for MNB. The tk is the spam term occurrence within an email and P(p) is classed as the prior probability. 1 and |V| are identified as the smoothing constant for the algorithm.

To test this algorithm, MNB module was loaded from the Scikit-learn library. The parameters for this model are optional. If none is specified, the default values are: Alpha value set to '1.0', Fit Prior is set to 'True' and Class Prior is set to 'None' .

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

The algorithm-1 shows the pseudocode for Multinomial Naïve Bayes with spam classification where ‘‘Tr’’ is Training and ‘‘Te’’ is Testing. The P(ˆ tk |p) is the estimating/predicting variable, also known as the conditional probability.

Analyze and Prediction:

In the actual dataset, we chose only 2 features :

Text: the text of the Emails data

2 Labels : Labels of Emails which can be either Spam or no spam

1: spam

0: no spam

Accuracy on test set:

We got a accuracy of 98.02% on test set.

Saving the Trained Model:

Once you’re confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or .pkl file using a library like `pickle` .

Make sure you have `pickle` installed in your environment.

Next, let’s import the module and dump the model into `.pkl` file

BIO-INSPIRED OPTIMIZATION ALGORITHMS

There are two bio-inspired optimization approaches that are discussed here which helped to improve the results of the experiments, i.e., Particle Swarm Optimization and Genetic Algorithm.

A. PARTICLE SWARM OPTIMIZATION ALGORITHM

The PSO is based on the swarming methods observed in fish or birds. The particles are evaluated based on their best position and overall global position. Particles within a search space are scattered to find the global best position.

The Py swarm’s library offers different calculations and techniques for PSO to be used with an ML model such as feature subset selection or parameter tuning optimization. As researched in the previous sections, the feature selection can reduce feature space but can also discard some features that can be useful during the classification. Therefore, PSO will be used to tune and find the hyper-parameter for a given ML/NN model.

B. GENETIC ALGORITHM

The GA algorithm is an evolutionary algorithm based on Darwinian natural selection that selects the fittest individual from the given population. This involves the principle of variation, inheritance and selection. The algorithm maintains a population size and the individuals have a unique number. (Chromosomes) that are binary represented Implementation of the GA was conducted with the help of TPOT library. The program selects the best parameters from a given dictionary of parameters. The TPOT classifier is then trained with cross validation. The parameters given to the TPOT are as follows:

Generation: Number of times the pipeline will conduct the optimization processes. The default value is 100. The program has set this parameter as ‘10’.

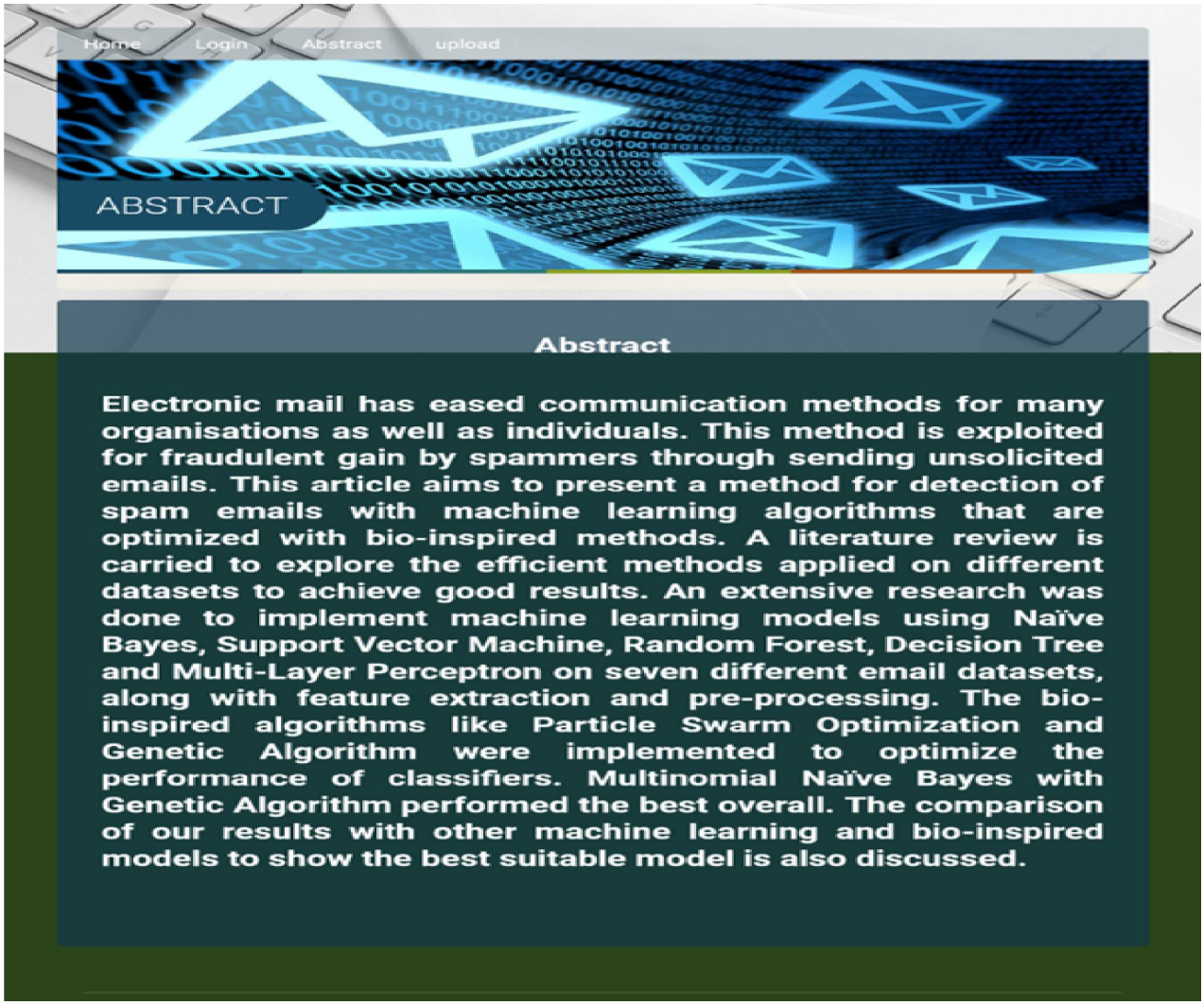
Population size: Number of individuals participating for Genetic programming within each generation. Default is 100. The program has set this parameter as ‘40’

Offspring size: Offspring to be produced in each generation. Default is 100. The program has set this parameter as ‘20’

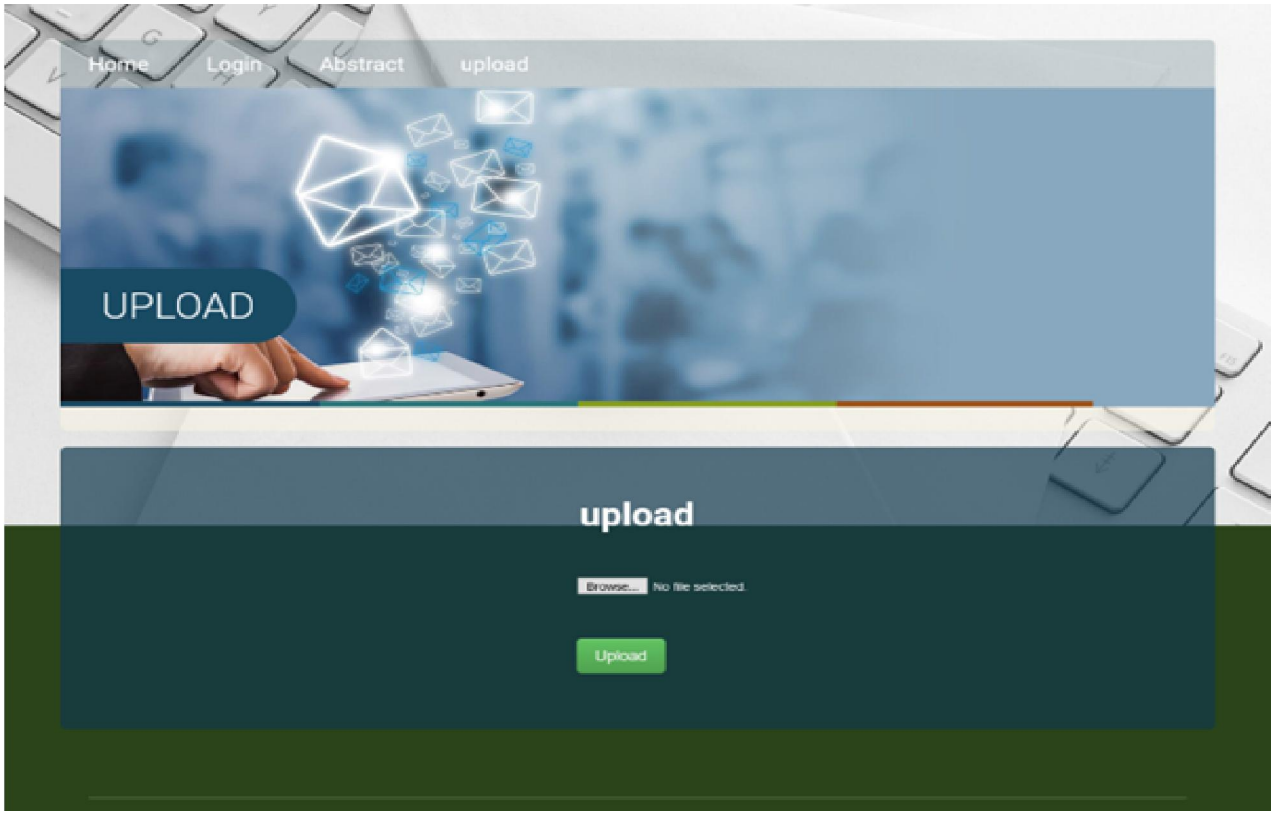
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

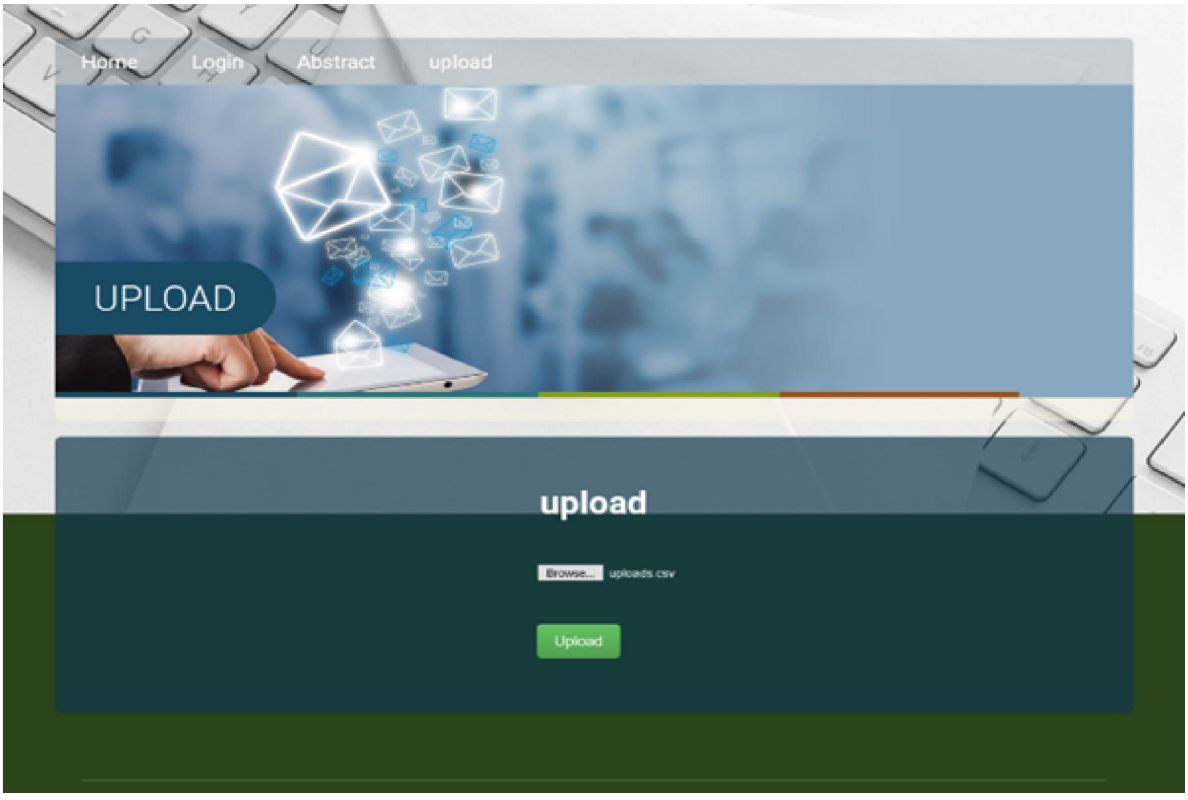
IX. OUTPUT

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023



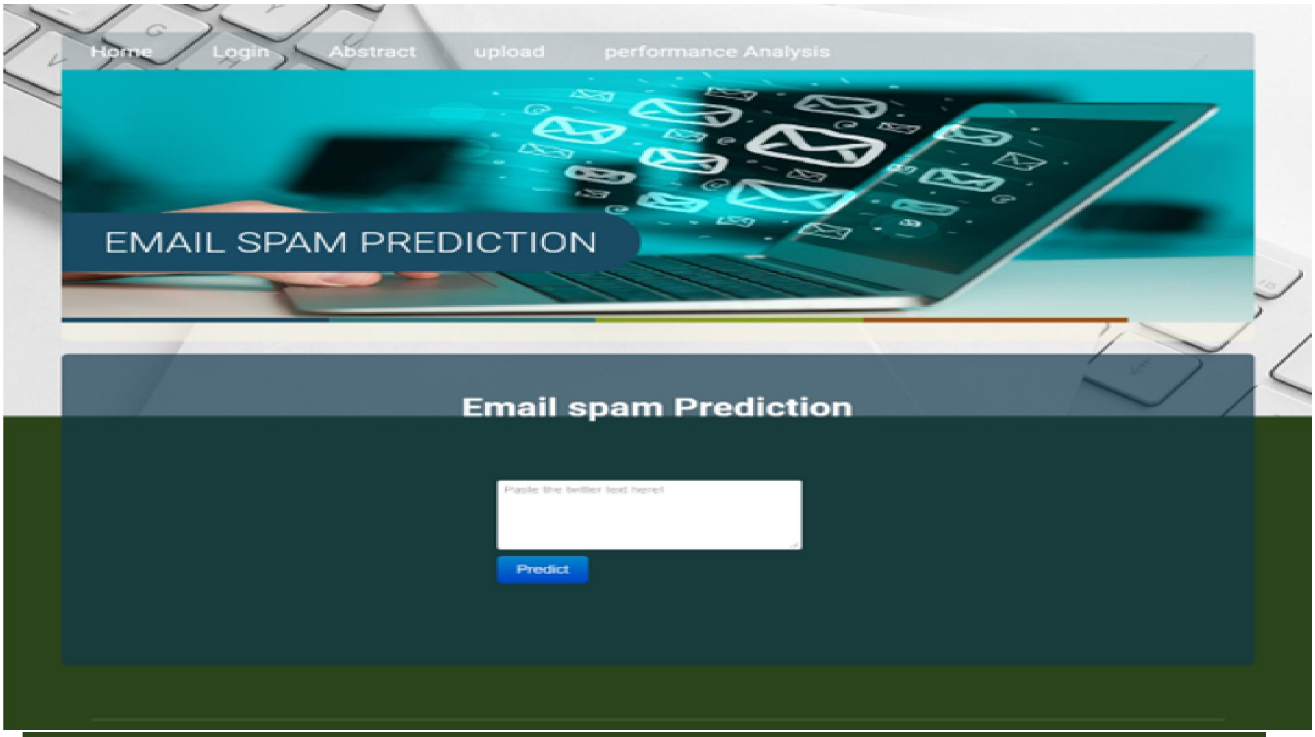
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

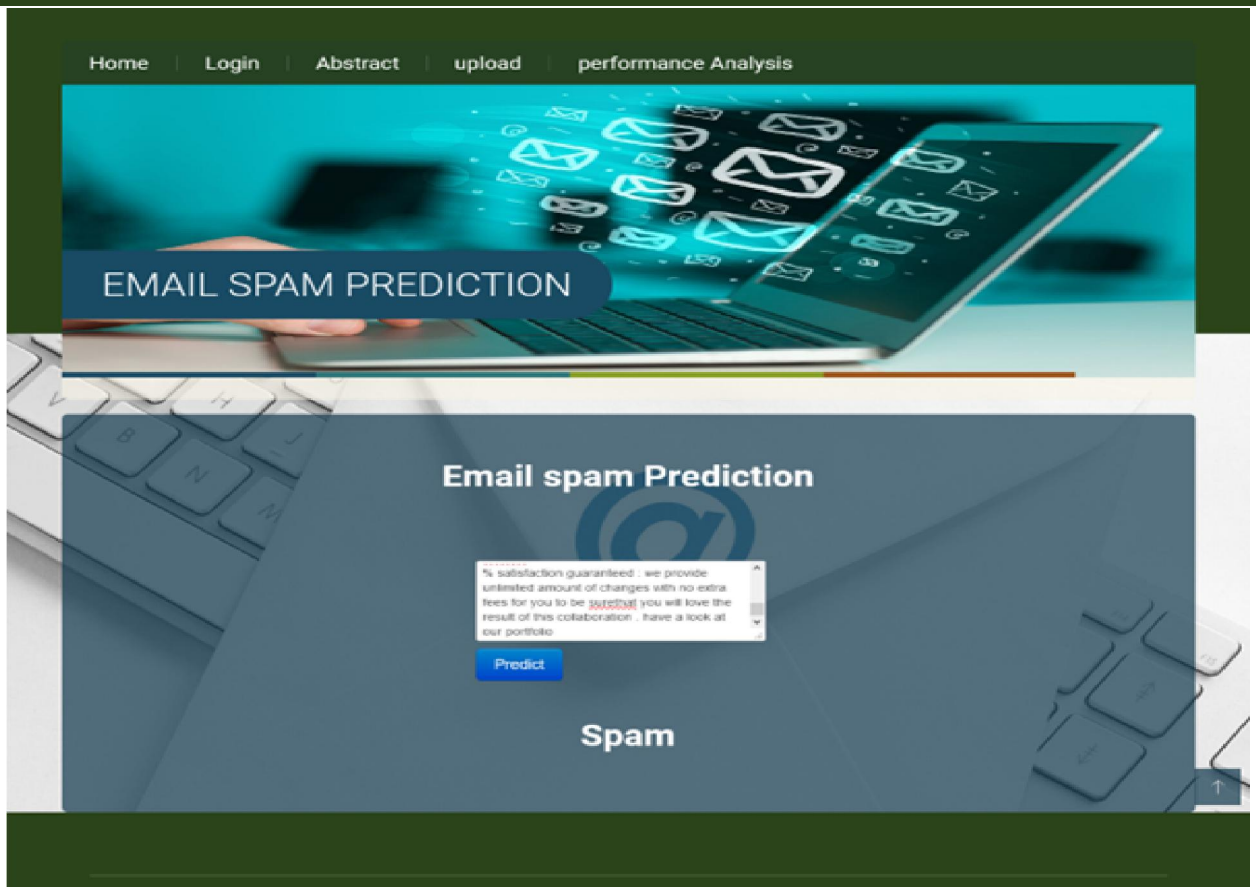




IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

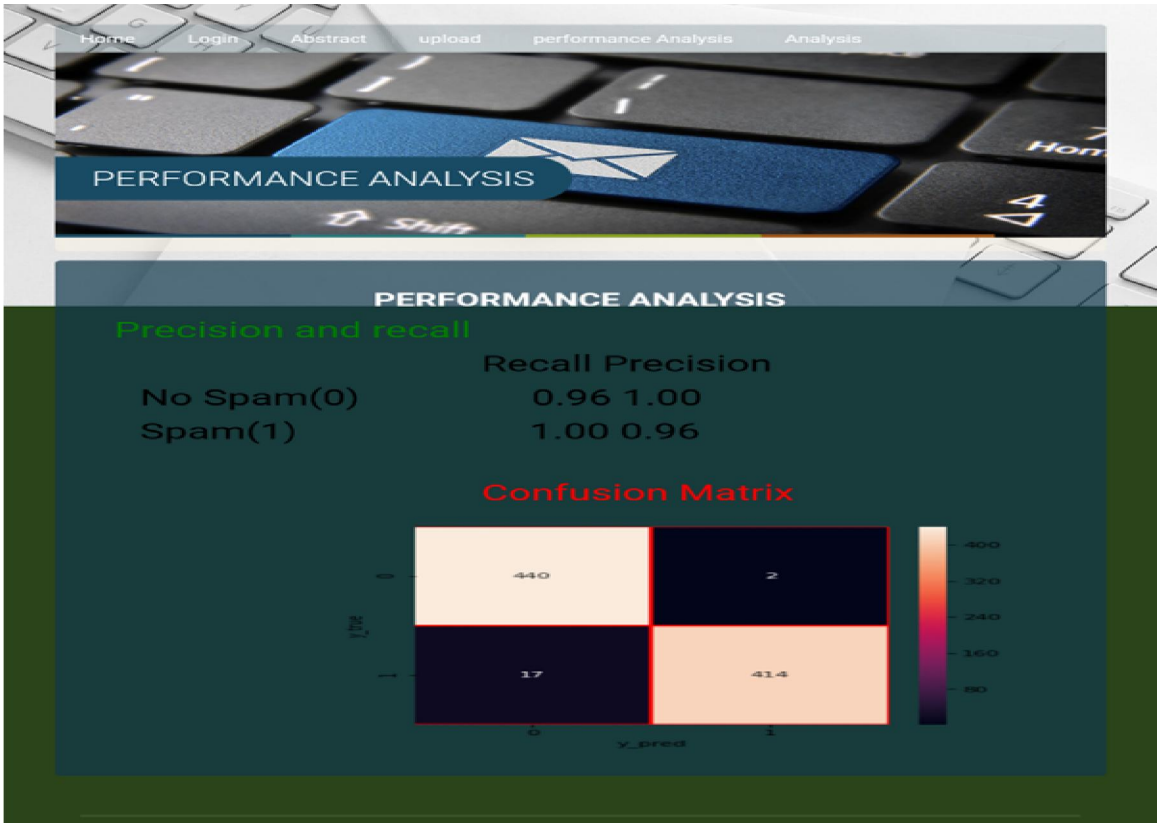
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

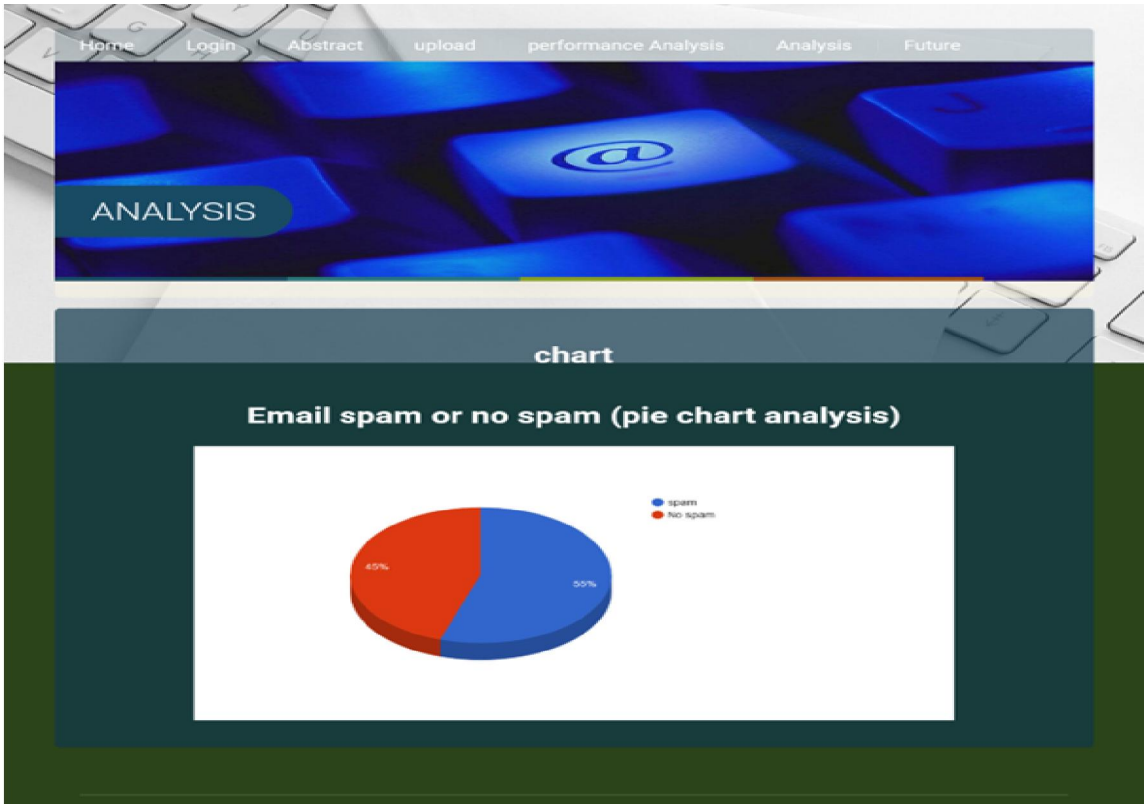




IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023





IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

IX. CONCLUSION

Spam detection and filtration gained the attention of a sizeable research community. The reason for a lot of research in this area is its costly and massive effect in many situations like consumer behavior and fake reviews.

The survey covers various machine learning techniques and models that the various researchers have proposed to detect and filter spam in emails and IoT platforms. The study categorized them as supervised, unsupervised, reinforcement learning, etc.

The study compares these approaches and provides a summary of learned lessons from each category. This study concludes that most of the proposed email and IoT spam detection methods are based on supervised machine learning techniques. A labeled dataset for the supervised model training is a crucial and time-consuming task.

Supervised learning algorithms SVM and Naïve Bayes outperform other models in spam detection. The study provides comprehensive insights of these algorithms and some future research directions for email spam detection and filtering.

REFERENCES

[1]. Emmanuel Gbenga Dada et al "Machine learning for email spam filtering: review, approaches and open research problems", Heliyon 5 (2019) e01802 Received 3 September 2018; Received in revised form 25 February 2019; Accepted 20 May 2019

[2]. Jai Batra et al "A comprehensive study of spam detection in e-mails using bio-inspired optimization techniques", International Journal of Information Management Data Insights 1 (2021) 100006 Received 20 October 2020; Received in revised form 30 November 2020; Accepted 19 December 2020

[3]. 1Dr. V. Malsoru, et al "Detecting Spam Email With Machine LearningOptimized With Bio-Inspired MetaheuristicAlgorithms", JAC : A Journal Of Composition Theory, ISSN : 0731-6755, Page No: 40-47

[4]. P. VANAJA, et al "Machine Learning based Optimization for Efficient Detection of Email Spam", Positif Journal, Issn No : 0048-4911, Page No : 310-319

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

[5]. K.Varun Kumar, et al "Machine Learning-based spam detection usingNaïve Bayes Classifier in comparison withLogistic Regression for improving accuracy", Journal of Pharmaceutical Negative Results ¦ Volume 13 ¦ Special Issue 4 ¦ 2022, Page No. 548-554

[6]. SIMRAN GIBSON, et al "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms, IEEE Access · January 2020, Page No. 187914-187932

[7]. N. Sutta, et al " A Study of Machine Learning Algorithms on Email Spam Classification", EPiC Series in Computing Volume69, 2020, Page No.170-179

[8]. Chode Abhinav, et al "Spam Mail Detection using Machine Learning", International Journal for Research in Applied Science & Engineering Technology (IJRASET)ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538, Page no.2327-2329

[9]. Thashina Sultana, et al "Email based Spam Detection", International Journal of Engineering Research & Technology (IJERT) IJERTV9IS060087 (Vol. 9 Issue 06, June-2020) Page No. 135-139

[10]. Rajesh Kumar J, et al "Email Spam Detection using Machine Learning Techniques", International Advanced Research Journal in Science, Engineering and Technology Vol. 8, Issue 6, June 2021, DOI: 10.17148/IARJSET.2021.8632 Page No. 189-193

[11]. Naresh Vinod Wankhade, et al "Paper on Spam Email Detection with Classification Using Machine Learning", INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY , IJIRT 156181, Page No. 1055-1059

[12]. Ms.A. Sowshna, et al "Detecting Spam Email With Machine Learning Optimized With Bio Inspired Metaheuristic Algorithms", International Journal of Scientific Development and Research (IJSDR), Page No. 160-163

[13]. Neha Karadkar, et al "Spam Mail Classification Using SVM and Genetic Algorithm", Journal of Emerging Technologies and Innovative Research (JETIR), Page No. e513-e517

[14]. Miss. Pratiksha Mantri, et al "A Proposed Paper on Spam Email Detection using Machine Learning", Journal of Emerging Technologies and Innovative Research (JETIR), Page No. a573-577

[15]. Nebojsa Bacanin, et al "Application of Natural Language Processing and Machine Learning Boosted with Swarm Intelligence for Spam Email Filtering", Mathematics 2022, 10, 4173. https://doi.org/10.3390/math10224173

[16]. Pooja Malhotra, et al "Spam Email Detection using Machine Learning and Deep Learning Techniques", https://ssrn.com/abstract=4145123

[17]. Omar Almomani, et al "A Hybrid Model Using Bio-Inspired Metaheuristic Algorithms for Network Intrusion Detection System", Tech Science Press, Page No. 410-429