iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44315

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 10 Issue VI June 2022- Available at www.ijraset.com

Spam Mail Detection using Machine Learning

Chode Abhinav[1], K Jayachandra[2], Kommu Pranith Kumar[3], V Sowmya[4]
Electronics and Computer Engineering Department, Sreenidhi Institute of Science and Technology, Hyderabad.

Abstract: ***Spam email is one of the most serious problems in the online world. Nowadays, a large portion of the population relies on available emails or communications from strangers. As a result, the fact that anyone can leave an email or a message opens the door for spammers to compose spam messages concerning our various interests. Spam fills up our inbox with unnecessary messages, slowing down our internet connection and stealing valuable information such as our contact information and accurate information. Detecting spammers and spam content is a major issue of research and time-consuming tasks. Email spam is when someone sends out a large number of emails in a short period of time. The purpose of spam filtering is to determine whether an email is spam or ham. With this proposed system the specified mail can be detected as spam or ham and also IP address of mail.***

Keywords: Spam or ham, Logistic Regression, IP address, Spam filtering, Machine Learning.

I. INTRODUCTION

One of the most successful and extensively utilized forms of communication is email. The appeal of email systems arises from the fact that they are inexpensive and quick to communicate with. Spam email is, unfortunately, a menace to email systems. Spam emails are unsolicited emails sent for the goal of making money by undesired users, generally known as spammers. The majority of email users' time is spent classifying spam emails. Multiple copies of the same communication are sent again and over, which not only costs the company money but also irritates the recipients. Spam emails not only intrude into users' inboxes, but they also generate a substantial volume of unnecessary data, reducing network capacity and utilization [1].

Many tests are conducted on spam to create algorithms capable of identifying spam. Email filtering is often classified based on the content related to their images, attachments, IP addresses or headers to provide data about the recipients. In this project, a proposed spam detection system (SMD) will identify email data into spam and ham [5].

II. LITERATURE REVIEW

On email spam detection, as well as social media and Twitter signaling spam detection, a lot of research and literature studies have been done. Because this is a relatively new area of research, there is no thorough systematic literature review on SMS spam detection. Although SMS communication first became popular in 2000, it gained traction in 2006 and acquired much greater traction after the introduction of Android phones. SMS spam is growing more popular with spammers as the number of individuals utilizing SMS as a way of communication grows. As a result, SMS spam detection research evolved out of need, and it primarily began around 2007. Our goal with this review is to gain proper knowledge in the field of spam mail detection, gain knowledge about the algorithms currently used for spam mail detection, their benefits and drawbacks, compare the accuracy of algorithms and identifying any gaps in current research so that need to be investigated further.

Spam and ham mails are classified using a variety of algorithms. On a spam base dataset, feature selection has a vital role in identifying the optimal classification method in terms of computational time, accuracy, misclassification rate, and precision, followed by algorithm selection.

III. METHODOLOGY

A. Machine Learning Algorithm

A machine learning method is employed in this project and adjusted to match the project's needs. This is due to the fact that machine learning algorithms excel at analyzing massive amounts of data. As the amount of data processed increases, it usually improves. This provides the system more practice and allows it to produce more accurate predictions. Machine learning allows for immediate adaptation. It detects new risks and responds appropriately. Because of its automated nature, it also saves time. There have been a lot of recent advances in spam detection using machine learning. Many various algorithms, such as Nave Bayes, Bayesian Nets, SVM, decision trees, random forests, and so on, have been tried and tested with very good accuracy results. As a result, methods based on machine learning are becoming increasingly popular. So this machine learning algorithm models are successful in spam detection and can be tested on specific or particular datasets.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

B. Logistic Regression Model

Logistic regression is a classification method that works well for binary classifications. This method could be great for detecting spam in our situation. This logistic model is used to estimate the probability of two-class response (like ham/spam) based on the given input values. Logistic Regression, often known as supervised machine learning, is one of the most widely used Machine Learning algorithms. It is method for predicting categorical dependent variables based on a set of independent variables. Logistic Regression is used for prediction of the output of a categorical dependent variable. As a result, the final value must be discrete or categorical. It must be yes or no, 0 or 1, true or false, and so on, but in case of giving actual numbers like 0 and 1, it gives probability values which falls in the between 0 and 1.

C. Data Loading and Understanding

Here data is loaded from the kaggle dataset then used pandas package to understand the dataset.Here an email dataset for the Spam mail detection system is taken. Various emails are selected at random from kaggle. For classification purposes, the dataset contains a total of 5572 emails, including both ham and spam emails. The dataset is divided into two sets in which one is for training and other for testing. Most of the time, training data is split into two sets: training and testing.In training model, we took 80 percent of dataset which is 4457 mails for training the dataset and we extracted the features by using spam filtering technique process. All the mails are separated into two sets X and Y then trained it by using algorithm.

In the testing model, we took 20 percent of dataset which is 1115 mails for training the dataset and we extracted the features by using spam filtering technique process and finally one set is considered as resultant output.Obtained results shows output either 0 or 1.

After the training then evaluated the model on the test data. The model had given accuracy for both training and testing data.

We can find IP address of a mail by using below link:

“https://whatismyipaddress.com/trace-email”

IV. RESULTS

```
model = LogisticRegression()
```

```
# training the Logistic Regression model with the training data
model.fit(X_train_features, Y_train)
```

```
LogisticRegression()
```

Fig 1. Training model

```
# prediction on training data

prediction_on_training_data = model.predict(X_train_features)
accuracy_on_training_data = accuracy_score(Y_train, prediction_on_training_data)
```

```
print('Accuracy on training data : ', accuracy_on_training_data)
```

```
Accuracy on training data :  0.9670181736594121
```

```
# prediction on test data

prediction_on_test_data = model.predict(X_test_features)
accuracy_on_test_data = accuracy_score(Y_test, prediction_on_test_data)
```

```
print('Accuracy on test data : ', accuracy_on_test_data)
```

```
Accuracy on test data :  0.9659192825112107
```

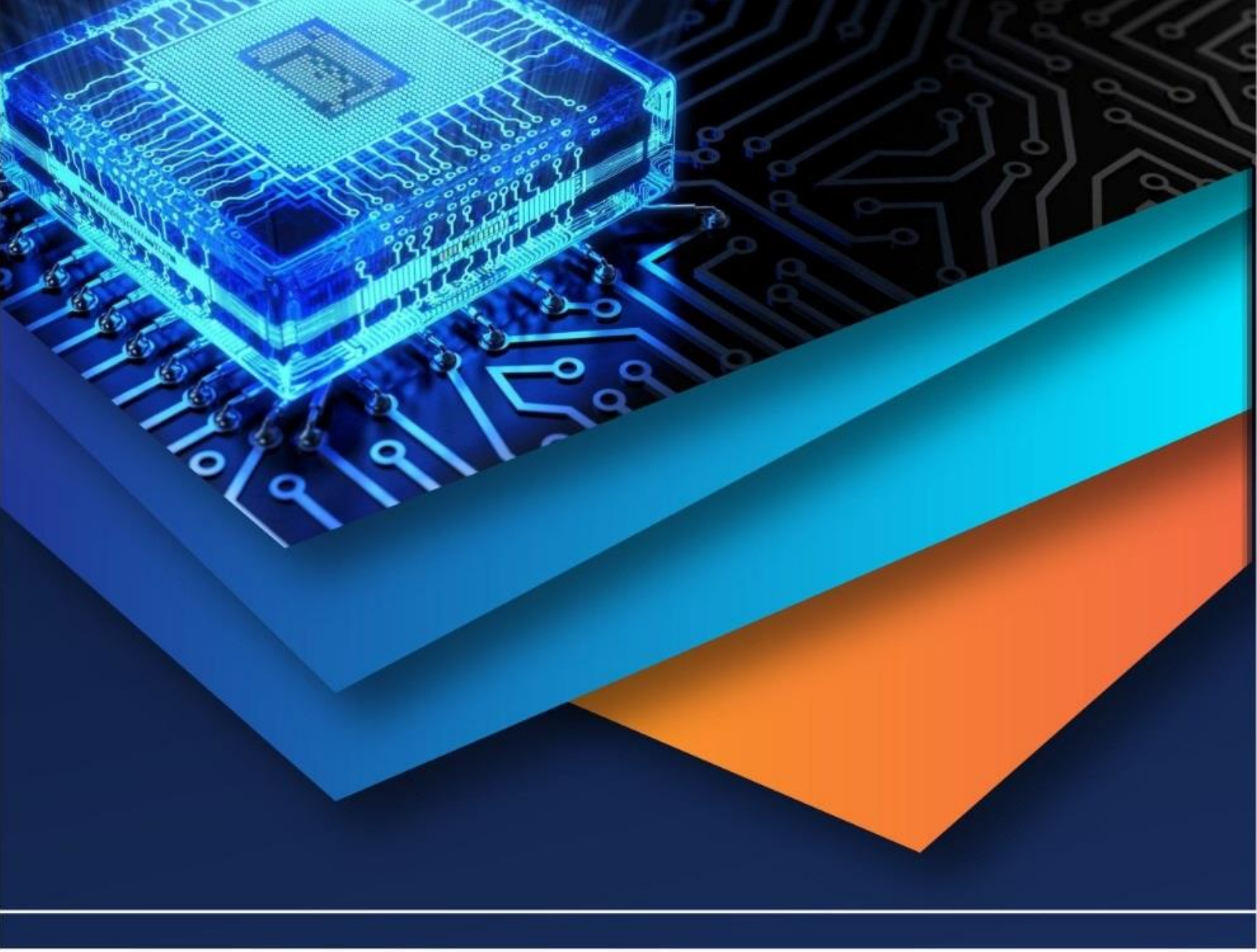
Fig 2. Evaluation of trained model

V. CONCLUSION

In today's age of communication and technology, spam email is one of the most demanding and unpleasant concerns on the internet. For safeguarding message and e-mail transmission, spam detection is very necessary. The accurate detection of spam is a big challenge, and researchers have proposed a lot of detection approaches. These approaches, are incapable of proper and efficient detection of spam. To solve this problem, we suggested a spam detection model based on machine learning prediction models.When compared to other current methods, the proposed method attained a high accuracy of 96 percent. As a result, the suggested system is structured in such a way that it recognises unsolicited and undesired mails and blocks them, hence minimising spam messages, which would be beneficial to people.

REFERENCES

[1] Machine Learning based Spam E-mail Detection, "https://www.researchgate.net/publication/326089990_Machine_Learning_based_Spam_Email_Detection".
[2] Machine Learning Techniques for Spam Detection in Email and IoT platforms: Analysis and Research Challenges, "https://doi.org/10.1155/2022/1862888".
[3] A Systematic Literature Review on SMS Spam Detection Techniques, "https://www.researchgate.net/publication/318298908".
[4] Onur Goker, "Spam filtering using Bigdata and Deeplearning" February 2018.
[5] Hybrid Machine Learning based E-mail Spam Filtering Technique
[6] "https://www.scribd.com/document/459203809/email-spam".
[7] Machine Learning Model Training,
[8] "https://r.search.yahoo.com/_ylt=Awrxx_15Doli1hMAnK67HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1653178106/RO=10/RU=https%3a%2f%2fblog.dominodatalab.com%2fwhat-is-machine-learning-model-training/RK=2/RS=dBbv3QtOGECSbSQQGdNAyISlE5c".
[9] A Spam Email Detection Mechanism for English Language Text Emails Using Deep Learning Approach, "https://www.researchgate.net/publication/348984746".



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY