NATIONAL INSTITUTE OF TECHNOLOGY

WARANGAL

# PROJECT ON : SPAM PREDICTION IN MAIL USING MACHINE LEARNING.

# PROJECT UNDER : DR.RAVI KUMAR JATOTH.

# PRESENTED BY : KRANTHI GOLLAPATI.

# ON THE OCCASION OF : SUMMER INTERNSHIP PROGRAM-2022.

# ROLL NO: 21MAC2R08.

# MATHEMATICS AND SCIENTIFIC COMPUTING (2021-2023).

CONTENTS:

1. ABSTRACT
2. INTRODUCTION
3. Spam
4. Spam Techniques
5. Types of Spam
6. How to fight with Spam
7. Machine Learning
8. Logistic Regression
9. Implementation
10. Conclusion

## Abstract:

* Email Spam has become a major problem nowadays, with Rapid growth of internet users, Email spams is also increasing. People are using them for illegal and unethical conducts, phishing and fraud. Sending malicious link through spam emails which can harm our system and can also seek in into your system. 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. So, it is needed to Identify those spam mails which are fraud, this project will identify those spam by using techniques of machine learning, this project will discuss the machine learning algorithms and apply all these algorithm on our data sets and best algorithm is selected for the email spam detection having best precision and accuracy..
* This is a project about unwanted email messages and inappropriate Usenet articles —what they are, who is sending them, how to stop them, and even how to outlaw them. This project about what has come to be called Internet spam\* This project for people who have seen their mailboxes fill up with useless messages and unsolicited advertisements. This project for people who are upset that they can't find the on-topic postings in their once-helpful Usenet newsgroups and fear that the community of newsgroup readers will dissolve in disgust. And it's a project for Internet service providers (ISPs) who are concerned about the growing toll that spam is taking on their systems—and are looking for ways to put an end to it once and for all.

Introduction:

Technology has become a vital part of life in today’s time. With each passing day, the use of the internet increases exponentially, and with it, the use of email for the purpose of exchanging information and communicating has also increased, it has become second nature to most people. While e-mails are necessary for everyone, they also come with unnecessary, undesirable bulk mails, which are also called Spam Mails [29]. Anyone with access to the internet can receive spam on their devices. Most spam emails divert people’s attention away from genuine and important emails and direct them towards detrimental situations. Spam emails are capable of filling up inboxes or storage capacities, deteriorating the speed of the internet to a great extent. These emails have the capability of corrupting one’s system by smuggling viruses into it, or steal useful information and scam gullible people. The identification of spam emails is a very tedious task and can get frustrating sometimes.

While spam detection can be done manually, filtering out a large number of spam emails can take very long and waste a lot of time. Hence, the need for spam detection softwares has become the need of the hour. To solve this problem, various spam detection techniques are used now. The most common technique for spam detection is the utilization of Naive Bayesian [5] method and feature sets that assess the presence of spam keywords. The main purpose is to demonstrate an alternative scheme, with the use of Neural Network (NN) [4] classification system that utilises a collection of emails sent by several users, is one of the objectives of this research. One other purpose is the development of spam detection with the help of Artificial Neural Networks, resulting in almost 98.8% accuracy.

Spam:

If you use email, it's likely that you've recently received a piece of spam—an unsolicited, unwanted message sent to you without your permission. Spam is the Internet's version of junk mail, telemarketing calls during dinner, crank phone calls, and leaflets pasted around town, all rolled up into a single annoying electronic bundle.

Spam is not democratic. If you are new to the Internet, you've probably seen only a few of these annoying messages. If you've been using the Internet for more than a few years, or if you participate in online discussion groups, you might receive a dozen or more of these messages each day. And if you administer a network for a business or university, you might be bombarded with hundreds.

Example:

Received: (from mail@localhost)

by apache.vineyard.net (8.8.5/8.8.5) id LAA01663

for ; Sat, 16 May 1998 11:57:57 -0400 (EDT)

From: [charles7713@yahoo.com](mailto:charles7713@yahoo.com)

Message-Id:

Received: from 209-142-2-72 . stk. inreach.net (209 . 142 .2 . 72 )

by apache.vineyard.net via smap/slg (VI. 3)

id sma001626; Sat May 16 11:57:27 1998

Date: Sat, 16 May 1998 05:18:34

To: Subject: Search Engines, 400 for 5.75 (1)

\*\*\* LIMITED TIME SPECIAL OFFER \*\*\*

For Only $5.75 (1) We Will Submit Your Web Site To Over 400 Of The Net's Hottest Search Engines, Directories & Indexes. If you're site isn't listed in the Search Engines, how can people find you to buy your products or services? • Your Competition Is Getting Noticed - Are You? Get Noticed By Your Prospects. Visit Our Web Site To Learn More: http: //www.tiffiny. com/sitesubmissions

Thank You (1)

The price for this service is $69 prepaid which covers the cost of submitting your site every three months for an entire year. We have shown the price of $5.75 to show you how inexpensive this program really is when the overall cost is annualized. Minimum 12 month term and full prepayment required.

Name removal requests.

Send to: TO: [webmaster@tiffiny.com](mailto:webmaster@tiffiny.com)

SUB : remove

This email from tiffiny.com has all the elements of a typical spam message:

• The message came from a business with which we had no prior relationship.

• It was sent from an email address (charles7713@yahoo.com) that either is ficti- tious or was created solely for the purpose of sending spam messages and has long since been discarded.

• The message advertises a service that is illegitimate, shady, or misleading at best. (The service being advertised is not S5.75, as the subject line says, but S5.75 per month, with a "minimum 12 month term and full prepayment required." Furthermore, there simply arent 400 "hot" search engines, directories, and indexes on the Internet.)

• The message does not clearly identify the person or group that has sent it.

• Removal requests sent to the address listed at the bottom, webmaster®tiffiny.com. were ignored.

• The company that's doing the advertising is not well known and typically isn't trying to establish a reputation or a loyal consumer following.

### **Account Deactivation Notices**

You and your coworkers likely have access to a countless number of accounts, both professionally and personally. If you suddenly received an email notification claiming an important account was deactivated, you’d probably check to see why it was deactivated. Hackers are counting on you to investigate by clicking the malicious link they provided in the email.

### **Compromised Credit Card**

If the hacker knows that you made a recent purchase, they could use that information to their advantage. Say you used your Chase card to buy a new TV. With that knowledge, the cybercriminal may send you a message that appears as if it came from Chase customer service. The email may also say something like, “Your card was compromised, please confirm credit details to protect your account.”

### **Tech Support Request**

If your company has an IT department, then you probably receive occasional emails from them regarding your IT. For example, they may notify you that your servers are going to be down for maintenance at a specific time. It’s easy to just go along with whatever they say since they are your technology experts. However, a hacker can easily mimic an internal email and request that you install “the latest update” for a program.

### **Fund Transfers**

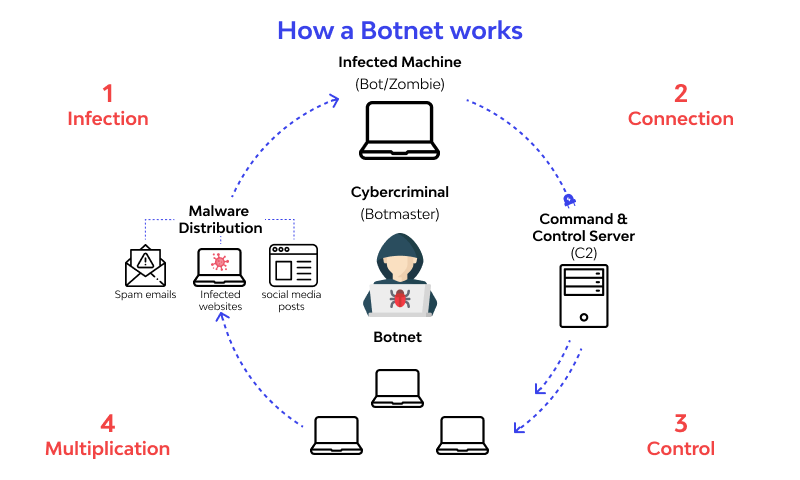
Imagine that your CEO or another executive is on travel and you receive an urgent email asking you to help them by transferring funds somewhere to secure a new partnership. What are you going to do? This type of phishing email is not only utilizing authority, but also telling you that the request is urgent. It’s not uncommon for a victim to panic in this scenario and send the money.

### **Fake Login Page**

### Sometimes it’s not the link in the email, but the destination. Some hackers create fake login pages and send phishing emails hoping to trick someone into logging into the fake website. The email may read something like, “We’ve updated our terms of service. Please log in to confirm your account.”

Spam Techniques:

Spammers use several different techniques to send spam, including the following:

* **Botnets.** Botnets enable spammers to use [command-and-control servers](https://www.techtarget.com/whatis/definition/command-and-control-server-CC-server) to both harvest email addresses and distribute spam.
* **Snowshoe spam.**With this technique, spammers use a wide range of [Internet Protocol (IP) addresses](https://www.techtarget.com/whatis/definition/IP-address) and email addresses with neutral reputations to distribute spam widely.
* **Blank email spam.**This technique involves sending email with an empty message body and subject line. It could be used in a directory harvest attack to validate email addresses by identifying invalid [bounced addresses](https://www.techtarget.com/whatis/definition/bounce-email-bounce-mail). In some instances, seemingly blank emails may hide [viruses](https://www.techtarget.com/searchsecurity/definition/virus) and [worms](https://www.techtarget.com/searchsecurity/definition/worm) that can spread through [Hypertext Markup Language](https://www.theserverside.com/definition/HTML-Hypertext-Markup-Language) code embedded in the email.
* **Image spam.**The message text, which is computer-generated and unintelligible to human readers, is stored as a [JPEG](https://www.techtarget.com/whatis/definition/JPEG-Joint-Photographic-Experts-Group) (Joint Photographic Experts Group) or GIF (Graphics Interchange Format) file and placed into the email body. This method attempts to avoid detection from text-based [spam filters](https://www.techtarget.com/searchsecurity/definition/spam-filter).

Types of Spam:

Email spam comes in many forms, depending on the purpose of the spammer, including the following:

* **Marketing messages.**This type of spam peddles unsolicited or illegal products or services.
* **Malware messages.**Some spam emails contain malware, which can trick users into divulging personal information, paying money or taking some action they would not normally do.
* **Frauds and scams.**The advance fee/Nigerian prince scam is a well-known example of email-based fraud. A user receives an email with an offer that purportedly results in a reward if they pay an advance fee or small deposit. Once they make the payment, the fraudster will invent further fees or just stop responding.
* **Antivirus warnings.**These messages "warn" a user about a virus infection and offer a "solution" to fix it. If the user takes the bait and clicks on a link in the email, the hacker can gain access to their system. The email may also download a malicious file to the device.
* **Sweepstakes winners.**Spammers send emails claiming that a recipient has won a sweepstakes or a prize. To collect the prize, the recipient must click on a link within the email. The link is malicious and is typically used to steal the user's personal information.

How to fight with spam:

Email spam filters, which may be part of a security application or an email system add-on, can catch many spam messages, depositing them in a user's spam folder rather than their inbox. However, it's impossible to completely eliminate spam. Some newer filters can read images and locate the text in them, but that may inadvertently filter out nonspam emails that contain images featuring text.

**1.Never give out or post your email address publicly:**  
You should remember that everyone can easily access the Internet. That means, spammers are also lurking on the Internet and are constantly seeking available email addresses which they will send spam emails to. Posting your email address publicly allows others to send spam emails to you, or worse, hack your account if you are using a weak password.

**2.Think before you click**  
There might be instances where your email service providers’ automated email filter mistakenly mark legitimate emails as spam email due to its content (e.g. the email contains a hyperlink). However, in most cases, emails marked as “SPAM” or redirected to the spam folder of your mailbox are sent by spammers. Subject of spam messages usually include offer of cheap prescription drugs, advertisements on new medicines, and status of packages from shipping companies. Make sure that you scrutinize the content of spam emails before opening any attachments (even if it looks like an innocent text or image file) or clicking on hyperlinks. Refrain from downloading contents blocked by your email service providers in such emails too.

**3.Do not reply to spam messages**  
Almost all spam messages are malicious emails sent by unknown sources. These sources could be hackers who aim to hack into the computers of their victims. Never respond to spam messages because through this, the spammer will know that the email address is active and thus, it increases the chance of your email to be constantly targeted by the spammer.

**4.Download spam filtering tools and anti-virus software**  
Spam filtering tools and anti-virus software can help to scan the emails that you received for malware. If the emails that you received contain malware, the malicious content would be quarantined and you would be prevented from opening it. This helps to alleviate the chance of emails containing malware from infecting your computer. As such, do select spam filtering tools and anti-virus software with such features to reduce your woes of having to decipher email contents.

**5.Avoid using your personal or business email address**  
Do not use your personal or business email address when registering in any online contest or service such as applications, deal updates, etc. Many spammers watch these groups or emailing lists to harvest new email addresses.

Machine Learning:

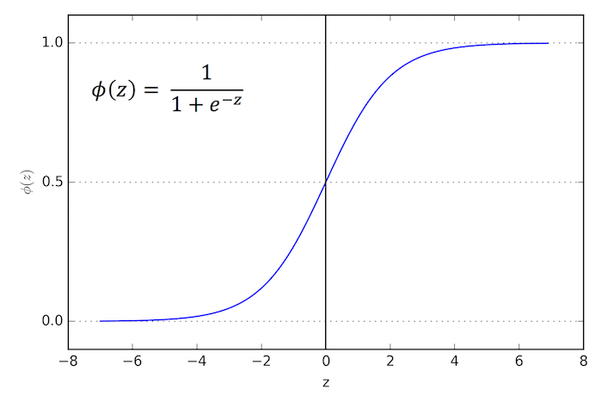
Machine learning (ML) is a type of artificial intelligence ([AI](https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence)) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning [algorithms](https://www.techtarget.com/whatis/definition/algorithm) use historical data as input to predict new output values.

Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies.

* **Supervised learning:** In this type of machine learning, [data scientists](https://www.techtarget.com/searchenterpriseai/definition/data-scientist) supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.
* **Unsupervised learning:** This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.
* **Semi-supervised learning:** This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled [training data](https://www.techtarget.com/searchenterpriseai/feature/Using-small-data-sets-for-machine-learning-models-sees-growth), but the model is free to explore the data on its own and develop its own understanding of the data set.
* **Reinforcement learning:**Data scientists typically use [reinforcement learning](https://www.techtarget.com/searchenterpriseai/definition/reinforcement-learning) to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.

## Logistic Regression:

Logistic Regression measures the relationship between the categorical dependent variable and one or more independent variables by **estimating** **probabilities** using a **logistic function**.From the definition it seems, the logistic function plays an important role in classification here but we need to understand what is logistic function and how does it help in estimating the probability of being in a class.



The formula mentioned in the above image is known as Logistic function or Sigmoid function and the curve called Sigmoid curve. The Sigmoid function gives an S shaped curve. The output of Sigmoid function tends towards 1 as z → ∞ and tends towards 0 as z → −∞. Hence Sigmoid/logistic function produces the value of dependent variable.which will always lie between [0,1] i.e the probability of being in a class.

Implementation:

#### **Step 1: Import Packages, Functions, and Classes**

First, you have to [import](https://realpython.com/absolute-vs-relative-python-imports/) Matplotlib for visualization and NumPy for array operations. You’ll also need LogisticRegression, train\_test\_split, TfidfVectorizer

and accuracy\_score from scikit-learn:

#importing libraries

import numpy as np

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score

#### **Step 2: Get Data**

In practice, you’ll usually have some data to work with. For the purpose of this project, let’s just read the data from the email data set using pandas module:

#loading data from csv file

data=pd.read\_csv('/content/mail\_data.csv')

print(data)

output:

|  | **Category** | **Message** |
| --- | --- | --- |
| **0** | ham | Go until jurong point, crazy.. Available only ... |
| **1** | ham | Ok lar... Joking wif u oni... |
| **2** | spam | Free entry in 2 a wkly comp to win FA Cup fina... |
| **3** | ham | U dun say so early hor... U c already then say... |
| **4** | ham | Nah I don't think he goes to usf, he lives aro... |
| **...** | ... | ... |
| **5567** | spam | This is the 2nd time we have tried 2 contact u... |
| **5568** | ham | Will ü b going to esplanade fr home? |
| **5569** | ham | Pity, \* was in mood for that. So...any other s... |
| **5570** | ham | The guy did some bitching but I acted like i'd... |
| **5571** | ham | Rofl. Its true to its name |

Step 3: in this step we are removing null values with strings

#replace null values with null strings

mail\_data=data.where((pd.notnull(data)),'')

mail\_data.head()

output:

| **Category** | **Message** |
| --- | --- |
| **0** | ham | Go until jurong point, crazy.. Available only ... |
| **1** | ham | Ok lar... Joking wif u oni... |
| **2** | spam | Free entry in 2 a wkly comp to win FA Cup fina... |
| **3** | ham | U dun say so early hor... U c already then say... |
| **4** | ham | Nah I don't think he goes to usf, he lives aro... |

Step4: we are checking number of rows and number of colums to separate the data

#checking the number of rows and columns

mail\_data.shape

output:

(5572, 2)

Step 5: we are giving true or false values in binary format

#label spam mail as 0 ,ham mail as 1

mail\_data.loc[mail\_data['Category']=='spam','Category',]=0

mail\_data.loc[mail\_data['Category']=='ham','Category',]=1

mail\_data.head()

output:

| **Category** | **Message** |
| --- | --- |
| **0** | 1 | Go until jurong point, crazy.. Available only ... |
| **1** | 1 | Ok lar... Joking wif u oni... |
| **2** | 0 | Free entry in 2 a wkly comp to win FA Cup fina... |
| **3** | 1 | U dun say so early hor... U c already then say... |
| **4** | 1 | Nah I don't think he goes to usf, he lives aro... |

Step6:separating X and Y columns

#seperating data into texts and label

X=mail\_data['Message']

Y=mail\_data['Category']

X

Output:

0 Go until jurong point, crazy.. Available only ...

1 Ok lar... Joking wif u oni...

2 Free entry in 2 a wkly comp to win FA Cup fina...

3 U dun say so early hor... U c already then say...

4 Nah I don't think he goes to usf, he lives aro...

...

5567 This is the 2nd time we have tried 2 contact u...

5568 Will ü b going to esplanade fr home?

5569 Pity, \* was in mood for that. So...any other s...

5570 The guy did some bitching but I acted like i'd...

5571 Rofl. Its true to its name

Name: Message, Length: 5572, dtype: object

Y

Output:

0 1

1 1

2 0

3 1

4 1

..

5567 0

5568 1

5569 1

5570 1

5571 1

Name: Category, Length: 5572, dtype: object

Step7:

#splittiing data into training data and testing data

X\_train,X\_test,Y\_train,Y\_test= train\_test\_split(X,Y,test\_size=0.2,random\_state=3)

print(X.shape)

print(X\_train.shape)

print(X\_test.shape)

* **random\_state** is an integer, an instance of numpy.RandomState, or None (default) that defines what pseudo-random number generator to use.
* **test\_size : *float or int, default=None***
* If float, should be between 0.0 and 1.0 and represent the proportion of the dataset to include in the test split. If int, represents the absolute number of test samples. If None, the value is set to the complement of the train size. If train\_size is also None, it will be set to 0.25.

output: (5572,)(4457,)

(1115,)

Step8:

#feature extraction

#transform the text data to feature vectors that cam be used as input to the logistic regression

feature\_extraction= TfidfVectorizer(min\_df=1,stop\_words='english',lowercase='True')

X\_train\_features= feature\_extraction.fit\_transform(X\_train)

X\_test\_features=feature\_extraction.transform(X\_test)

#convert Y\_train,Y\_test values as integers

Y\_train=Y\_train.astype('int')

Y\_test=Y\_test.astype('int')

print(X\_train)

Output:

3075 Don know. I did't msg him recently.

1787 Do you know why god created gap between your f...

1614 Thnx dude. u guys out 2nite?

4304 Yup i'm free...

3266 44 7732584351, Do you want a New Nokia 3510i c...

...

789 5 Free Top Polyphonic Tones call 087018728737,...

968 What do u want when i come back?.a beautiful n...

1667 Guess who spent all last night phasing in and ...

3321 Eh sorry leh... I din c ur msg. Not sad alread...

1688 Free Top ringtone -sub to weekly ringtone-get ...

Name: Message, Length: 4457, dtype: object

Step 9:

print(X\_train\_features)

output:

(0, 5413) 0.6198254967574347

(0, 4456) 0.4168658090846482

(0, 2224) 0.413103377943378

(0, 3811) 0.34780165336891333

(0, 2329) 0.38783870336935383

(1, 4080) 0.18880584110891163

(1, 3185) 0.29694482957694585

(1, 3325) 0.31610586766078863

(1, 2957) 0.3398297002864083

(1, 2746) 0.3398297002864083

(1, 918) 0.22871581159877646

(1, 1839) 0.2784903590561455

(1, 2758) 0.3226407885943799

(1, 2956) 0.33036995955537024

(1, 1991) 0.33036995955537024

(1, 3046) 0.2503712792613518

(1, 3811) 0.17419952275504033

(2, 407) 0.509272536051008

(2, 3156) 0.4107239318312698

(2, 2404) 0.45287711070606745

(2, 6601) 0.6056811524587518

(3, 2870) 0.5864269879324768

(3, 7414) 0.8100020912469564

(4, 50) 0.23633754072626942

(4, 5497) 0.15743785051118356

: :

(4454, 4602) 0.2669765732445391

(4454, 3142) 0.32014451677763156

(4455, 2247) 0.37052851863170466

(4455, 2469) 0.35441545511837946

(4455, 5646) 0.33545678464631296

(4455, 6810) 0.29731757715898277

(4455, 6091) 0.23103841516927642

(4455, 7113) 0.30536590342067704

(4455, 3872) 0.3108911491788658

(4455, 4715) 0.30714144758811196

(4455, 6916) 0.19636985317119715

(4455, 3922) 0.31287563163368587

(4455, 4456) 0.24920025316220423

(4456, 141) 0.292943737785358

(4456, 647) 0.30133182431707617

(4456, 6311) 0.30133182431707617

(4456, 5569) 0.4619395404299172

(4456, 6028) 0.21034888000987115

(4456, 7154) 0.24083218452280053

(4456, 7150) 0.3677554681447669

(4456, 6249) 0.17573831794959716

(4456, 6307) 0.2752760476857975

(4456, 334) 0.2220077711654938

(4456, 5778) 0.16243064490100795

(4456, 2870) 0.31523196273113385

Step 10:

#training the model

#logistic regression

model=LogisticRegression()

#training the logistic regression model with the training data

model.fit(X\_train\_features,Y\_train)

output:

LogisticRegression()

Step 11:

#evaluting the trrained model

#prediction on training data

y\_pred\_train=model.predict(X\_train\_features)

y\_accuracy\_train=accuracy\_score(Y\_train,y\_pred)

print('accuracy on training data : ',y\_accuracy\_train)

output:

accuracy on training data : 0.9670181736594121

#prediction on test data

y\_pred\_test=model.predict(X\_test\_features)

y\_accuracy\_test=accuracy\_score(Y\_test,y\_pred\_test)

print('accuracy on test data:',y\_accuracy\_test)

output:

accuracy on test data: 0.9659192825112107

step 11:

input\_mail=['']

#convert text to feature vector

input\_data\_features=feature\_extraction.transform(input\_mail)

#making predictt

prediction=model.predict(input\_data\_features)

print(prediction)

if prediction[0]==1:

  print('it is an ham mail')

else:

  print('spam mail')

output:

[1]

it is an ham mail

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. Besides advertising, these may contain links to phishing or malware hosting websites set up to steal confidential information. Spam is a serious issue that is not just annoying to the end-users but also financially damaging and a security risk. 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.

References:

* <https://ieeexplore.ieee.org/document/9183098>
* stopping Spam by .Alan Schwartz and Simson Garfinkel Copyright © 1998 O'Reilly & Associates. Inc. All rights reserved. Printed in the United States of America. Published by O'Reilly & Associates. Inc.. 101 Morris Street. Sebastopol. CA 95472.
* <https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML#:~:text=Machine%20learning%20(ML)%20is%20a,to%20predict%20new%20output%20values>.
* 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