E-MAIL SYSTEM (GMAIL CLONE) WITH CUSTOM SPAM CLASSIFIER

UCS663: CONVERSATIONAL AI
DATA SCIENCE
MINI PROJECT

Submitted By:

Tushar Mittal

3CS10

101916042

Submitted To:

Dr. Sahil Sharma

Assistant Professor

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

May 2022

ABSTRACT

The aim of this project is to build an electronic mail system, i.e. basically a gmail clone with backend programmed in Django Framework, frontend in HTML, CSS and Javascript and database used being SQLite. The system has a mail client server that performs the function of sending and receiving mails, replying to received mails and archiving mails. Further it has been pre-trained to classify mails as spam or ham (using the Naïve-Bayes Classifier). Anyone can comfortably use the system by registering as a new user in a few steps or by logging in using the registered credentials. Viewing the mails as spam or ham is also very easy since it shows whether a received or sent mail in the respective inbox is spam or ham directly.

TABLE OF CONTENTS

INTRODUCTION

In today’s globalized world, email is a primary source of communication. This communication can vary from personal, business, corporate to government. With the rapid increase in email usage, there has also been increase in the SPAM emails. SPAM emails, also known as junk email involves nearly identical messages sent to numerous recipients by email. Apart from being annoying, spam emails can also pose a security threat to computer system. It is estimated that spam cost businesses on the order of $100 billion in 2007.

In this project, I aim to perform automatic spam filtering in my self-created email system. I have tried to identify patterns using Naïve Bayes classification algorithm hence classifying the received emails as HAM or SPAM.

Naive Bayes classifiers are a popular statistical technique of e-mail filtering. They typically use bag-of-words features to identify spam e-mail, an approach commonly used in text classification. Naive Bayes spam filtering is a baseline technique for dealing with spam that can tailor itself to the email needs of individual users and give low false positive spam detection rates that are generally acceptable to users. It is one of the oldest ways of doing spam filtering, with roots in the 1990s.

The Naïve Bayes classification algorithm works very well in this scenario giving an accuracy of 97%.

RELATED WORK

Email is such an integral part of our day-to-day life that it has become the primary source of communication for most. It is easy to use, safe and very fast. With billions of users worldwide and more than 300 billion emails being exchanged every day, email has come a long way since it was invented in 1971.

Today, people all over the world have a host of email clients to choose from. According to the latest data, as of January 2022, the two most used email clients in the world are Apple and Gmail. At 57.16 percent, Apple has the majority of the email client market share. It's used by more than half of the world's email users to send and receive emails. In the race to capture the majority of market share, companies like Apple and Google work continuously for the betterment of their respective email systems. Classifying Emails as spam or ham is a major task they need to deal.

For Example: Google's email client Gmail spam filter works in the following way:

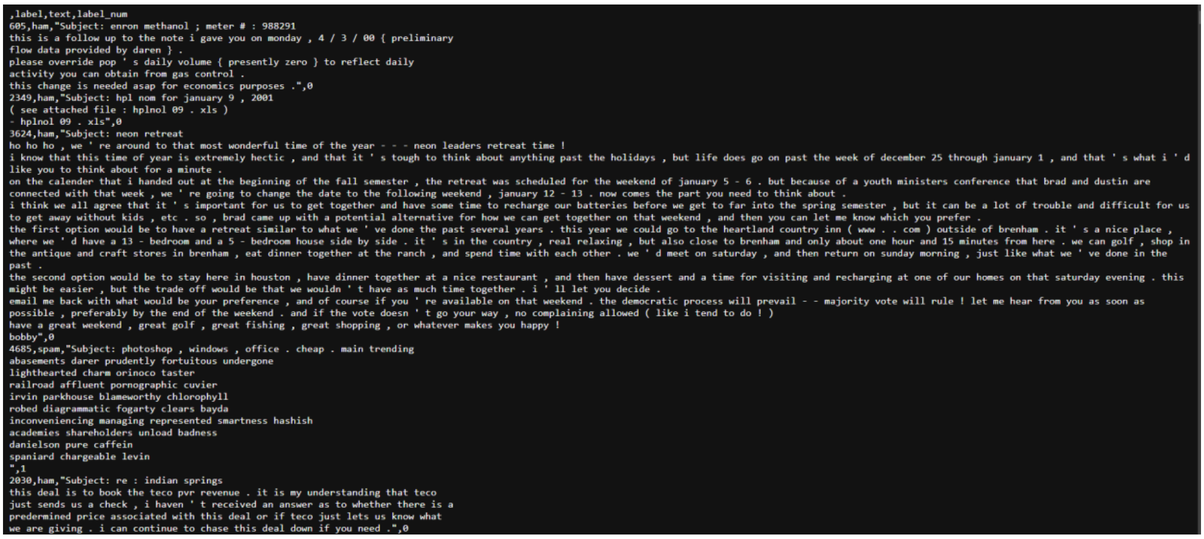
1. First, it checks the email of the sender against Gmail's database of blacklisted domains.
2. If the email passes that (if the email or domain is unknown), Gmail will then check any links against its database of known malicious links and compare them to links in the incoming email.
3. After this, Gmail will also check for spelling and grammatical errors and go through its list of trigger words that are heavily featured in known spam emails.
4. Gmail also uses an in-house machine learning framework called **Tensorflow** – alongside some smart AI – to train new spam filters moving forward.

And Apple's Iphone works in the following way to filter spam:
Traditionally, machine learning techniques formalize a problem of clustering of spam message collection through the objective function. The objective function is a maximization of similarity between messages in clusters, which is defined by k-nearest neighbor (kNN) algorithm.
Unfortunately, above approach do not provide good enough performance to filter spam e-mails for iPhone. Thus, Apple applies artificial bee-based decision tree (ABBDT) to filter spam e-mails for iPhone. In the proposed approach, decision tree is used to filter spam e-mails. In addition, artificial bee algorithm is used to ameliorate the testing accuracy of decision tree.

DATASET

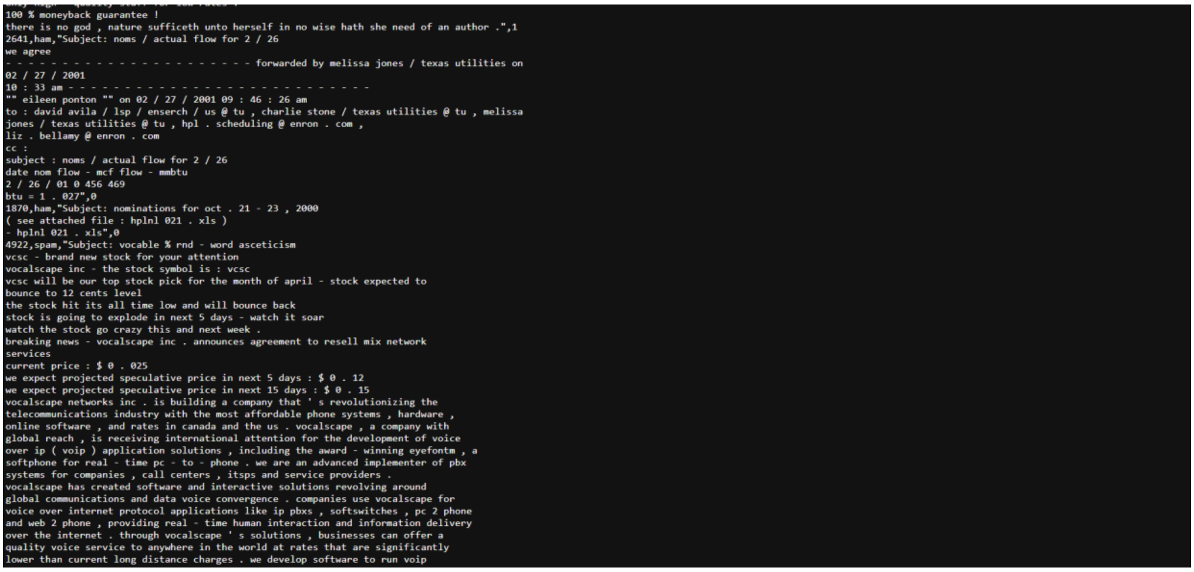
As data is the foundation of any model, selecting the correct dataset is the most critical stage in any machine learning task. I have chosen the following dataset for training the model by Naïve Bayes Classification Algorithm from the net and uploaded it is a part in my Git repository:

https://github.com/tush7301/Mail_system/blob/main/mail/spam_ham_dataset.csv



```
,label,text,label_num
605,ham,"Subject: enron methanol ; meter # : 988291
this is a follow up to the note i gave you on monday , 4 / 3 / 00 { preliminary
flow data provided by daren } .
please override pop ' s daily volume { presently zero } to reflect daily
activity you can obtain from gas control .
this change is needed asap for economics purposes .",0
2349,ham,"Subject: hpl nom for january 9 , 2001
( see attached file : hplnol 09 . xls )
- hplnol 09 . xls",0
3624,ham,"Subject: neon retreat
ho ho ho , we ' re around to that most wonderful time of the year - - - neon leaders retreat time !
i know that this time of year is extremely hectic , and that it ' s tough to think about anything past the holidays , but life does go on past the week of december 25 through january 1 , and that ' s what i ' d
like you to think about for a minute .
on the calender that i handed out at the beginning of the fall semester , the retreat was scheduled for the weekend of january 5 - 6 . but because of a youth ministers conference that brad and dustin are
connected with that week , we ' re going to change the date to the following weekend , january 12 - 13 . now comes the part you need to think about .
i think we all agree that it ' s important for us to get together and have some time to recharge our batteries before we get to far into the spring semester , but it can be a lot of trouble and difficult for us
to get away without kids , etc . so , brad came up with a potential alternative for how we can get together on that weekend , and then you can let me know which you prefer .
the first option would be to have a retreat similar to what we ' ve done the past several years . this year we could go to the heartland country inn ( www . . com ) outside of brenham . it ' s a nice place ,
where we ' d have a 13 - bedroom and a 5 - bedroom house side by side . it ' s in the country , real relaxing , but also close to brenham and only about one hour and 15 minutes from here . we can golf , shop in
the antique and craft stores in brenham , eat dinner together at the ranch , and spend time with each other . we ' d meet on saturday , and then return on sunday morning , just like what we ' ve done in the
past .
the second option would be to stay here in houston , have dinner together at a nice restaurant , and then have dessert and a time for visiting and recharging at one of our homes on that saturday evening . this
might be easier , but the trade off would be that we wouldn ' t have as much time together . i ' ll let you decide .
email me back with what would be your preference , and of course if you ' re available on that weekend . the democratic process will prevail - - majority vote will rule ! let me hear from you as soon as
possible , preferably by the end of the weekend . and if the vote doesn ' t go your way , no complaining allowed ( like i tend to do ! )
have a great weekend , great golf , great fishing , great shopping , or whatever makes you happy !
bobby",0
4685,spam,"Subject: photoshop , windows , office . cheap . main trending
abasements darer prudently fortuitous undergone
lighthearted charm orinoco taster
railroad affluent pornographic cuvier
irvin parkhouse blameworthy chlorophyll
robed diagrammatic fogarty clears bayda
inconveniencing managing represented smartness hashish
academies shareholders unload badness
danielson pure caffein
spaniard chargeable levin
",1
2030,ham,"Subject: re : indian springs
this deal is to book the teco pvr revenue . it is my understanding that teco
just sends us a check , i haven ' t received an answer as to whether there is a
predermined price associated with this deal or if teco just lets us know what
we are giving . i can continue to chase this deal down if you need .",0
```

FIGURE 1: A SECTION OF THE DATASET USED FOR TRAINING THE CLASSIFIER



```
100 % moneyback guarantee !
there is no god , nature sufficeth unto herself in no wise hath she need of an author .",1
2641,ham,"Subject: noms / actual flow for 2 / 26
we agree
- - - - - - - - - - - - - - - - - - - - - - forwarded by melissa jones / texas utilities on
02 / 27 / 2001
10 : 33 am - - - - - - - - - - - - - - - - - - - - - - - - - - -
"" eileen ponton "" on 02 / 27 / 2001 09 : 46 : 26 am
to : david avila / lsp / enserch / us @ tu , charlie stone / texas utilities @ tu , melissa
jones / texas utilities @ tu , hpl . scheduling @ enron . com ,
liz . bellamy @ enron . com
cc :
subject : noms / actual flow for 2 / 26
date nom flow - mcf flow - mmbtu
2 / 26 / 01 0 456 469
btu = 1 . 027",0
1870,ham,"Subject: nominations for oct . 21 - 23 , 2000
( see attached file : hplnl 021 . xls )
- hplnl 021 . xls",0
4922,spam,"Subject: vocable % rnd - word asceticism
vcsc - brand new stock for your attention
vocalscape inc - the stock symbol is : vcsc
vcsc will be our top stock pick for the month of april - stock expected to
bounce to 12 cents level
the stock hit its all time low and will bounce back
stock is going to explode in next 5 days - watch it soar
watch the stock go crazy this and next week .
breaking news - vocalscape inc . announces agreement to resell mix network
services
current price : $ 0 . 025
we expect projected speculative price in next 5 days : $ 0 . 12
we expect projected speculative price in next 15 days : $ 0 . 15
vocalscape networks inc . is building a company that ' s revolutionizing the
telecommunications industry with the most affordable phone systems , hardware ,
online software , and rates in canada and the us . vocalscape , a company with
global reach , is receiving international attention for the development of voice
over ip ( voip ) application solutions , including the award - winning eyefontm , a
softphone for real - time pc - to - phone . we are an advanced implementer of pbx
systems for companies , call centers , itsps and service providers .
vocalscape has created software and interactive solutions revolving around
global communications and data voice convergence . companies use vocalscape for
voice over internet protocol applications like ip pbxs , softswitches , pc 2 phone
and web 2 phone , providing real - time human interaction and information delivery
over the internet . through vocalscape ' s solutions , businesses can offer a
quality voice service to anywhere in the world at rates that are significantly
lower than current long distance charges . we develop software to run voip
```

FIGURE 2: A SECTION OF THE DATASET USED FOR TRAINING THE CLASSIFIER

METHODOLOGY

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable. Bayes’ theorem states the following relationship, given class variable y and dependent feature vector x_1 through x_n, :

$$P(y \mid x_1, \dots, x_n) = \frac{P(y) P(x_1, \dots, x_n \mid y)}{P(x_1, \dots, x_n)}$$

Using the naive conditional independence assumption that

$$P(x_i | y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i | y),$$

for all i, this relationship is simplified to:

$$P(y \mid x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \dots, x_n)}$$

In spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. They require a small amount of training data to estimate the necessary parameters.

Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

Multinomial Naïve Bayes:

MultinomialNB implements the naive Bayes algorithm for multinomially distributed data, and is one of the two classic naive Bayes variants used in text classification (where the data are typically represented as word vector counts, although tf-idf vectors are also known to work well in practice). The distribution is parametrized by vectors $\theta_y=(\theta_{y1},\dots,\theta_{yn})$ for each class y, where n is the number of features (in text classification, the size of the vocabulary) and θ_{yi} is the probability $P(x_i|y)$ of feature i appearing in a sample belonging to class y.

The parameters θ_y is estimated by a smoothed version of maximum likelihood, i.e. relative frequency counting:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

where $Ny_i = \sum_x \in Tx_i$ is the number of times feature i appears in a sample of class y in the training set T, and $N_y = \sum_i N_{yi}$ is the total count of all features for class y.

The smoothing priors $\alpha \geq 0$ accounts for features not present in the learning samples and prevents zero probabilities in further computations. Setting $\alpha = 1$ is called Laplace smoothing, while $\alpha < 1$ is called Lidstone smoothing.

The email system has been pretrained using Multinomial Naïve bayes classification algorithm of the sklearn library. The algorithm works quite well with the used dataset and gives an overall accuracy of more than 97% , which is clearly seen on the mails it classifies as ham and spam in the inbox of users.

PROPOSED SOLUTION ARCHITECTURE

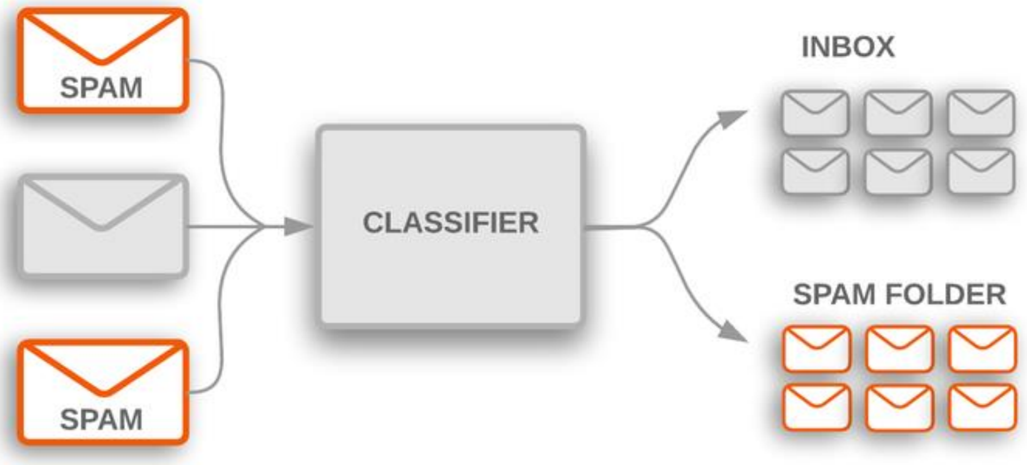


FIGURE 3: BASIC ARCHITECTURE

When a message is received or sent via the proposed email system, it undergoes a series of steps on the backend.

1. The message is first preprocessed and tokenized.
2. Then the tokens undergo probability estimation.
3. Then the machine applies Multinomial Naïve bayes classifier to classify the messages as spam or ham.
4. If the message is classified as spam, then the message is tagged spam and moved to a separate inbox and if the message is classified as legitimate, it is tagged ham and is processed as usual.

(Naive Bayes classifiers work by correlating the use of tokens (typically words, or sometimes other things), with spam and non-spam e-mails and then using Bayes' theorem to calculate a probability that an email is or is not spam.)

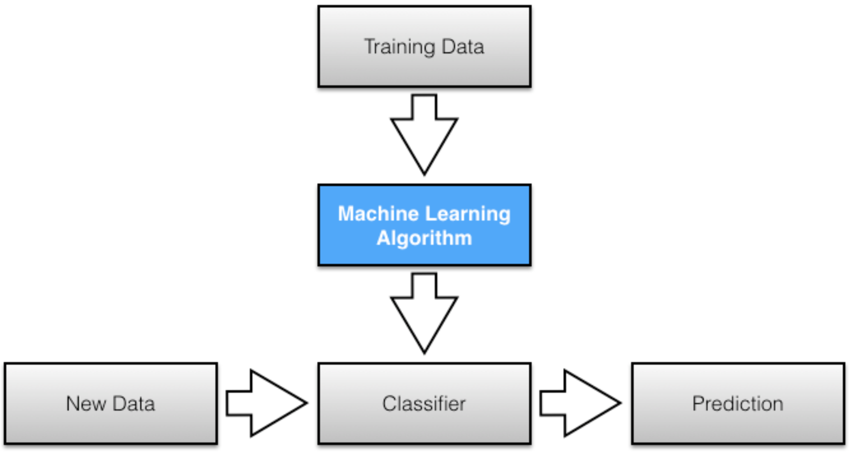


FIGURE 4: GENERAL WORKING OF A CLASSIFIER

The steps can be visualized in the following way:

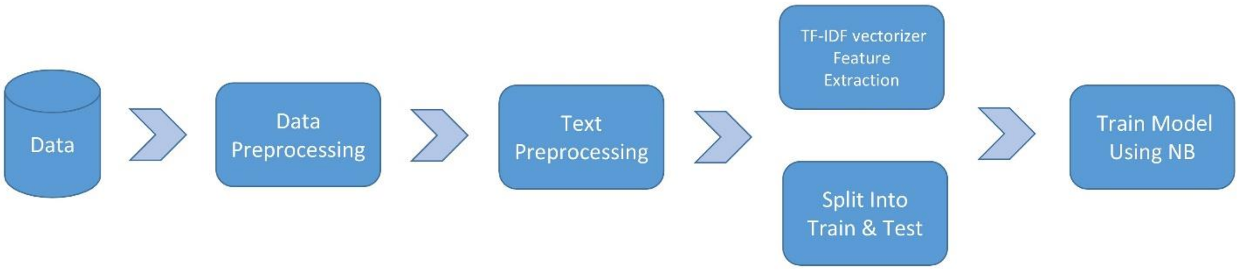


FIGURE 5: STEPS TO TRAIN THE MODEL

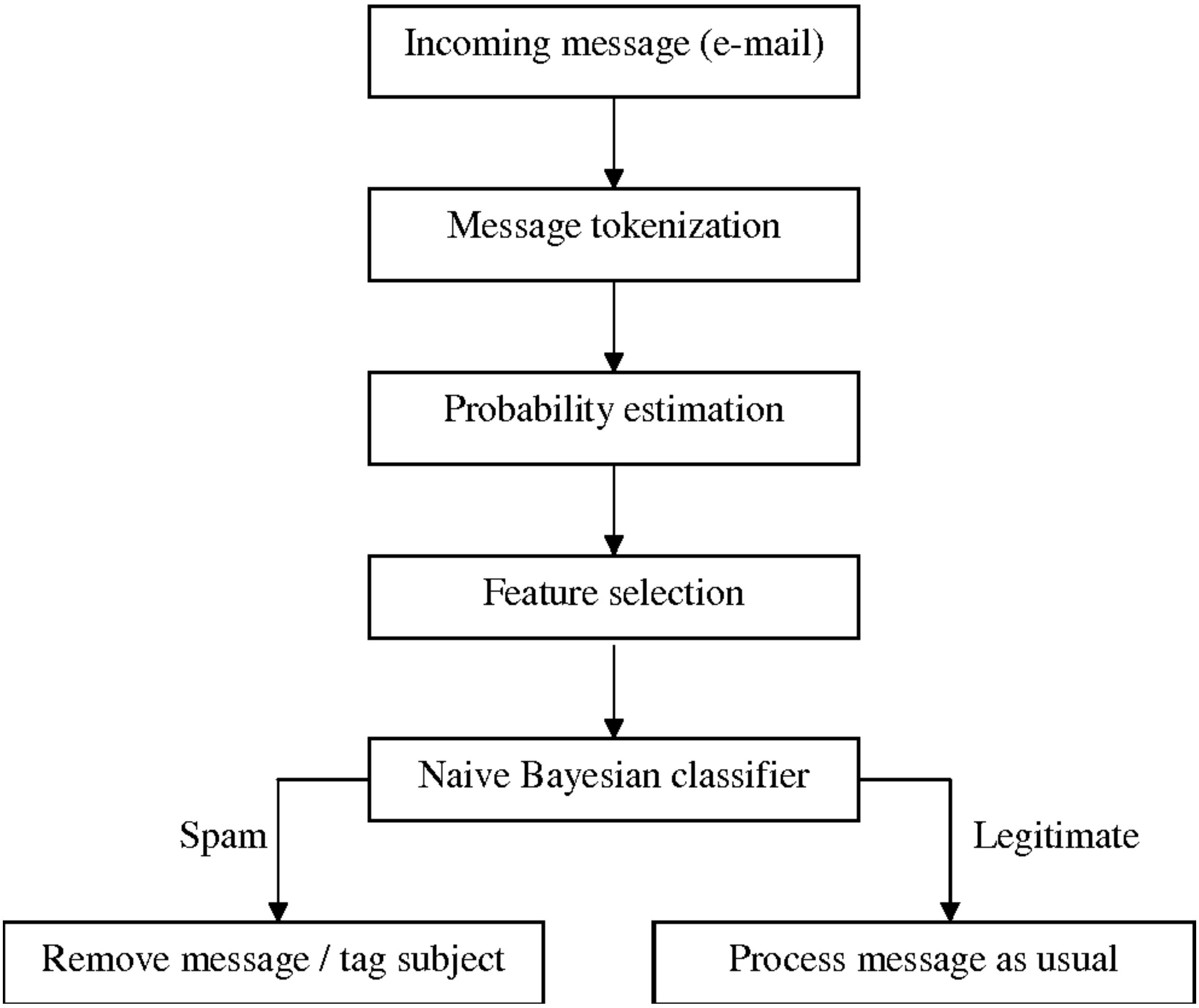
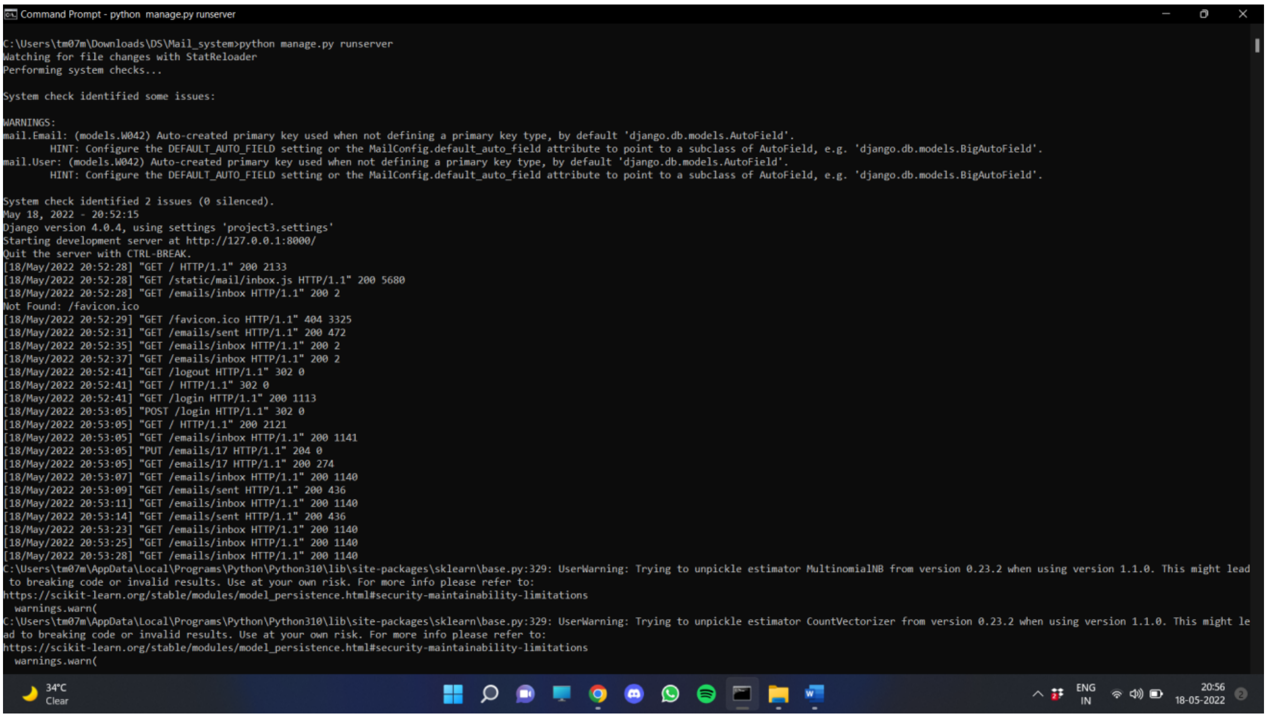


FIGURE 6: HOW AN INCOMING MAIL IS PROCESSED AS SPAM

EXPERIMENTATION AND RESULTS



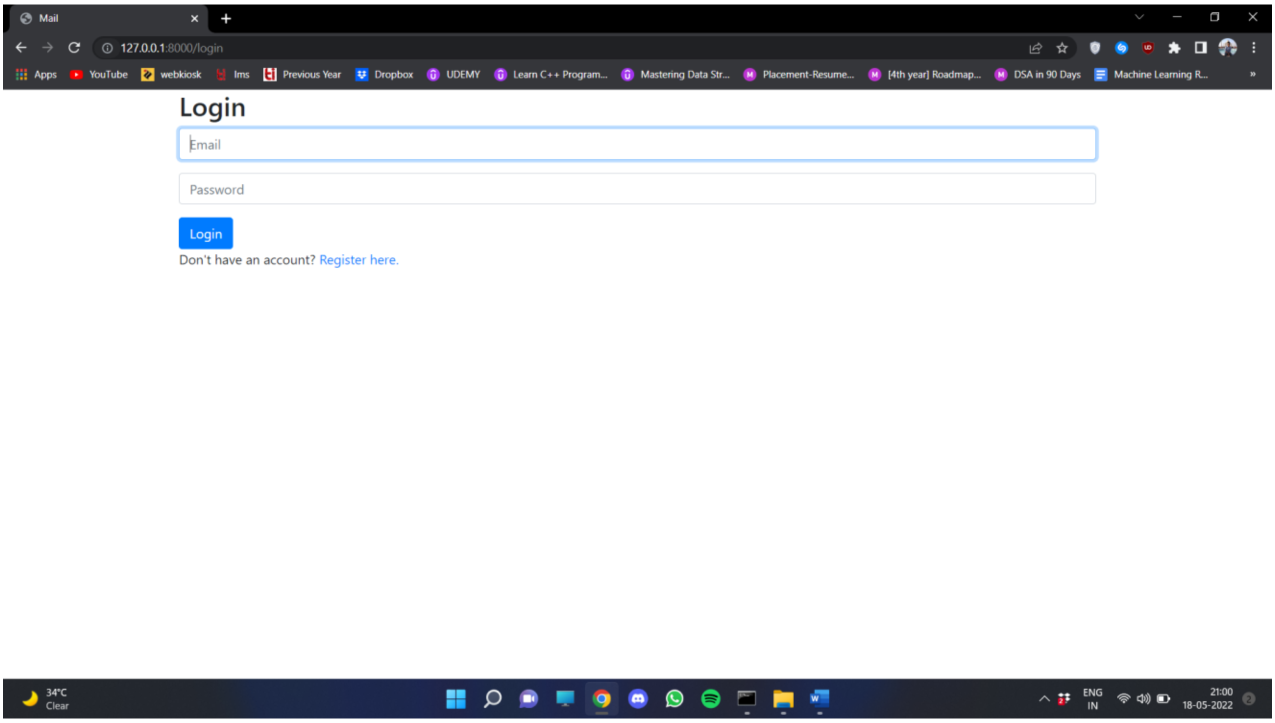
```
C:\Users\tm07m\Downloads\DS\Mail_system>python manage.py runserver
Watching for file changes with StatReloader
Performing system checks...

System check identified some issues:

WARNINGS:
mail.Email: (models.W042) Auto-created primary key used when not defining a primary key type, by default 'django.db.models.AutoField'.
        HINT: Configure the DEFAULT_AUTO_FIELD setting or the MailConfig.default_auto_field attribute to point to a subclass of AutoField, e.g. 'django.db.models.BigAutoField'.
mail.User: (models.W042) Auto-created primary key used when not defining a primary key type, by default 'django.db.models.AutoField'.
        HINT: Configure the DEFAULT_AUTO_FIELD setting or the MailConfig.default_auto_field attribute to point to a subclass of AutoField, e.g. 'django.db.models.BigAutoField'.

System check identified 2 issues (0 silenced).
May 18, 2022 - 20:52:15
Django version 4.0.4, using settings 'project3.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
[18/May/2022 20:52:28] "GET / HTTP/1.1" 200 2133
[18/May/2022 20:52:28] "GET /static/mail/inbox.js HTTP/1.1" 200 5680
[18/May/2022 20:52:28] "GET /emails/inbox HTTP/1.1" 200 2
Not Found: /favicon.ico
[18/May/2022 20:52:29] "GET /favicon.ico HTTP/1.1" 404 3325
[18/May/2022 20:52:31] "GET /emails/sent HTTP/1.1" 200 472
[18/May/2022 20:52:35] "GET /emails/inbox HTTP/1.1" 200 2
[18/May/2022 20:52:37] "GET /emails/inbox HTTP/1.1" 200 2
[18/May/2022 20:52:41] "GET /logout HTTP/1.1" 302 0
[18/May/2022 20:52:41] "GET / HTTP/1.1" 302 0
[18/May/2022 20:52:41] "GET /login HTTP/1.1" 200 1113
[18/May/2022 20:53:05] "POST /login HTTP/1.1" 302 0
[18/May/2022 20:53:05] "GET / HTTP/1.1" 200 2121
[18/May/2022 20:53:05] "GET /emails/inbox HTTP/1.1" 200 1141
[18/May/2022 20:53:05] "PUT /emails/17 HTTP/1.1" 204 0
[18/May/2022 20:53:05] "GET /emails/17 HTTP/1.1" 200 274
[18/May/2022 20:53:07] "GET /emails/inbox HTTP/1.1" 200 1140
[18/May/2022 20:53:09] "GET /emails/sent HTTP/1.1" 200 436
[18/May/2022 20:53:11] "GET /emails/inbox HTTP/1.1" 200 1140
[18/May/2022 20:53:14] "GET /emails/sent HTTP/1.1" 200 436
[18/May/2022 20:53:23] "GET /emails/inbox HTTP/1.1" 200 1140
[18/May/2022 20:53:25] "GET /emails/inbox HTTP/1.1" 200 1140
[18/May/2022 20:53:28] "GET /emails/inbox HTTP/1.1" 200 1140
C:\Users\tm07m\AppData\Local\Programs\Python\Python310\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator MultinomialNB from version 0.23.2 when using version 1.1.0. This might lead
 to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/modules/model_persistence.html#security-maintainability-limitations
  warnings.warn(
C:\Users\tm07m\AppData\Local\Programs\Python\Python310\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator CountVectorizer from version 0.23.2 when using version 1.1.0. This might le
ad to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/modules/model_persistence.html#security-maintainability-limitations
  warnings.warn(
```

FIGURE 7: RUNNING THE SERVER THROUGH COMMAND PROMPT TO ACCESS THE TERMINAL



Login

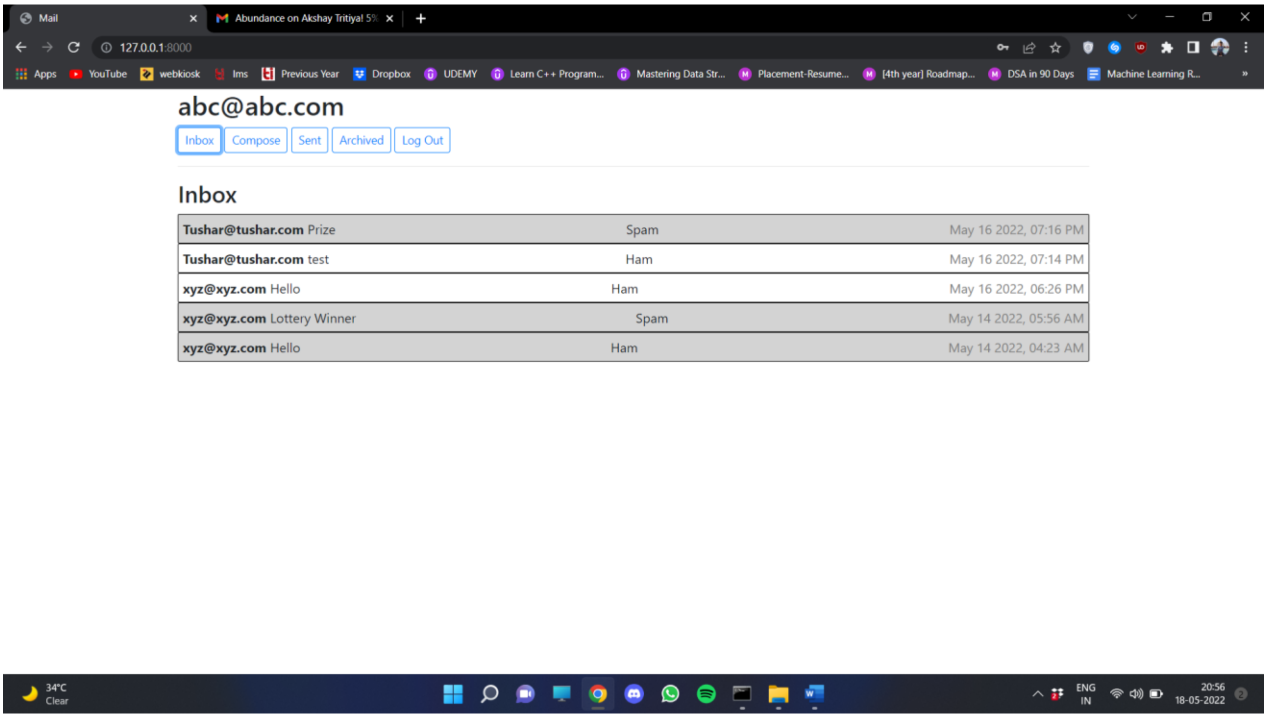
Email

Password

Login

Don't have an account? Register here.

FIGURE 8: LOGIN PAGE



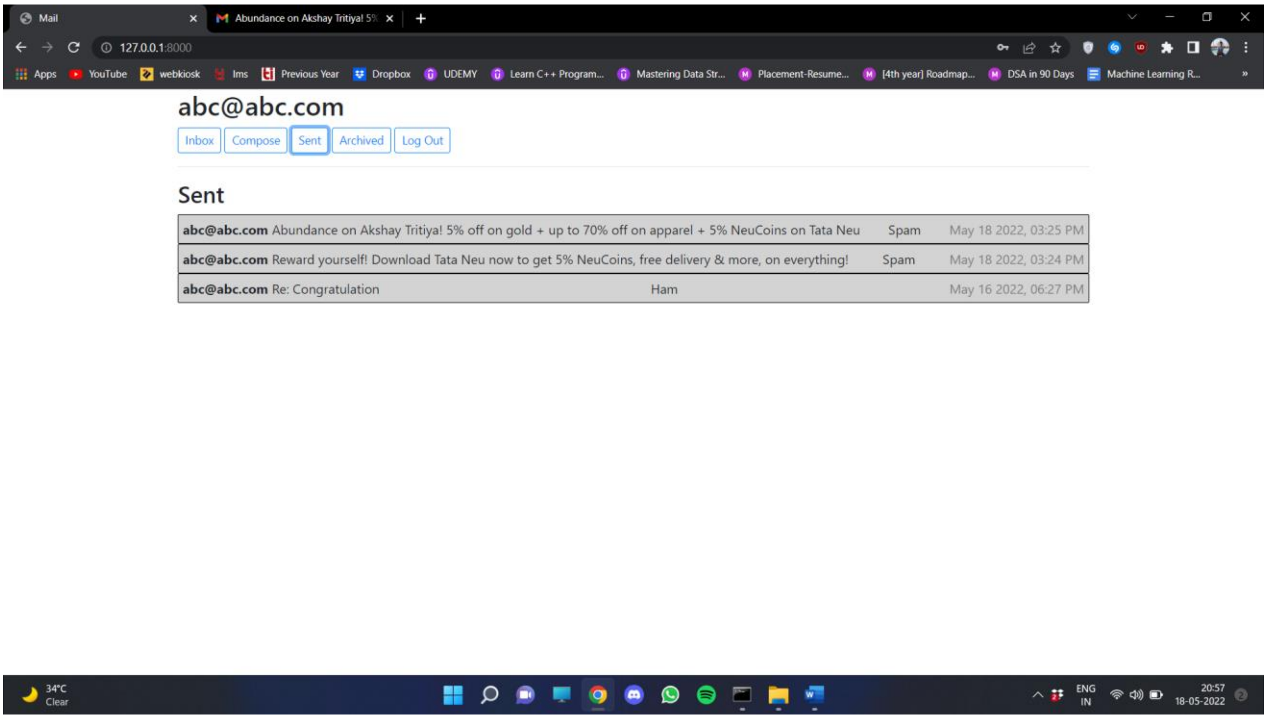
abc@abc.com

Inbox | Compose | Sent | Archived | Log Out

Inbox

Tushar@tushar.com Prize	Spam	May 16 2022, 07:16 PM
Tushar@tushar.com test	Ham	May 16 2022, 07:14 PM
xyz@xyz.com Hello	Ham	May 16 2022, 06:26 PM
xyz@xyz.com Lottery Winner	Spam	May 14 2022, 05:56 AM
xyz@xyz.com Hello	Ham	May 14 2022, 04:23 AM

FIGURE 9: SAMPLE USER 1-INBOX



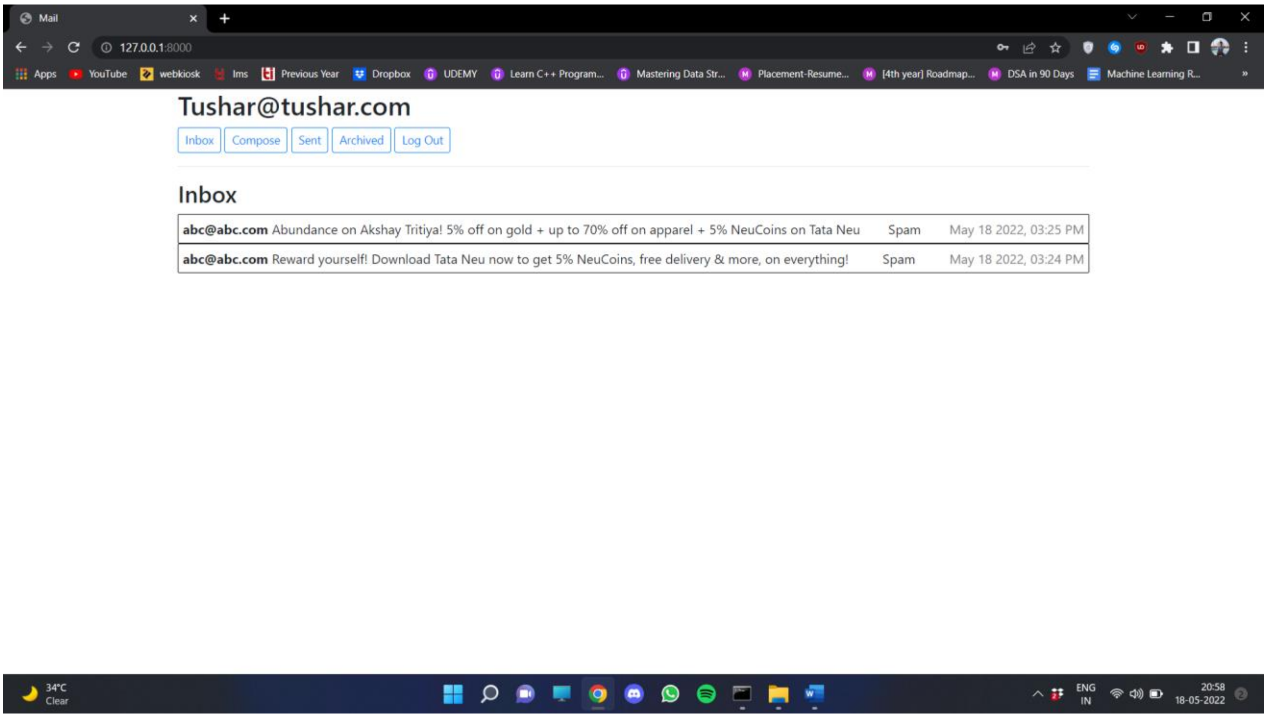
abc@abc.com

Inbox | Compose | Sent | Archived | Log Out

Sent

abc@abc.com Abundance on Akshay Tritiya! 5% off on gold + up to 70% off on apparel + 5% NeuCoins on Tata Neu	Spam	May 18 2022, 03:25 PM
abc@abc.com Reward yourself! Download Tata Neu now to get 5% NeuCoins, free delivery & more, on everything!	Spam	May 18 2022, 03:24 PM
abc@abc.com Re: Congratulation	Ham	May 16 2022, 06:27 PM

FIGURE 10: SAMPLE USER 1-SENT

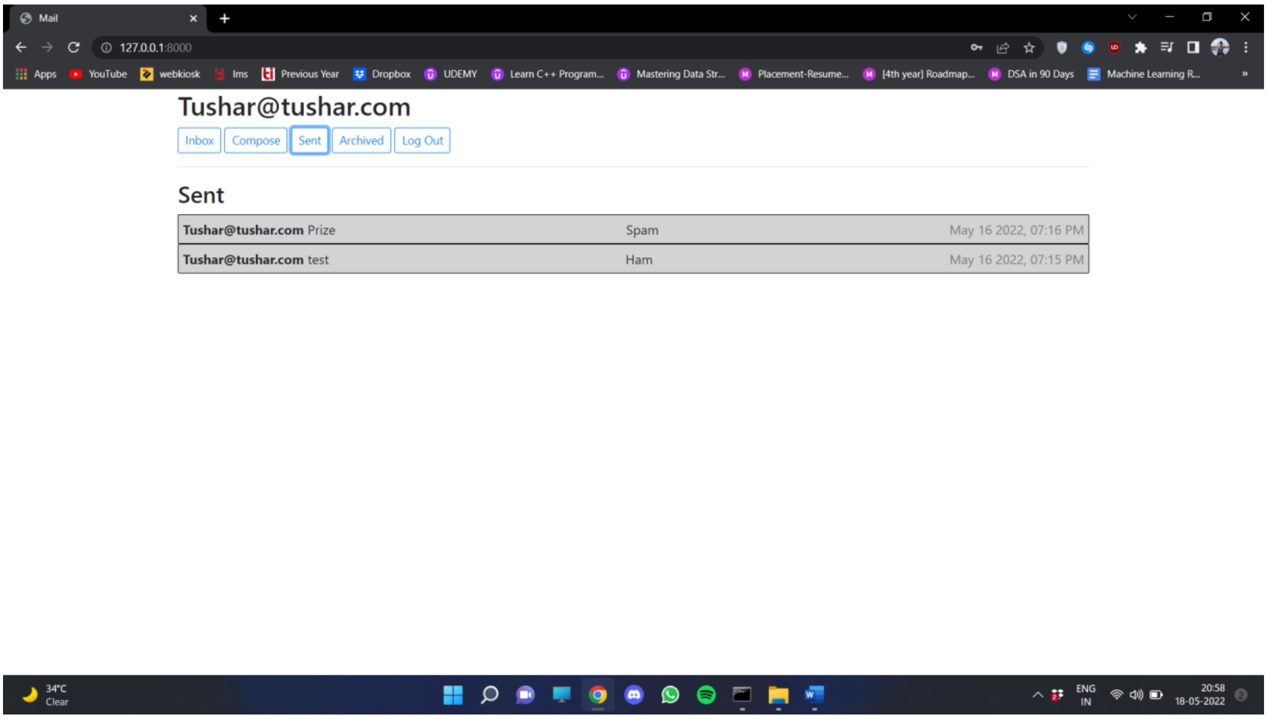


Tushar@tushar.com

Inbox

abc@abc.com Abundance on Akshay Tritiya! 5% off on gold + up to 70% off on apparel + 5% NeuCoins on Tata Neu	Spam	May 18 2022, 03:25 PM
abc@abc.com Reward yourself! Download Tata Neu now to get 5% NeuCoins, free delivery & more, on everything!	Spam	May 18 2022, 03:24 PM

FIGURE 11: SAMPLE USER 2-INBOX



Tushar@tushar.com

Sent

Tushar@tushar.com Prize	Spam	May 16 2022, 07:16 PM
Tushar@tushar.com test	Ham	May 16 2022, 07:15 PM

FIGURE 12: SAMPLE USER 2-SENT

CONCLUSION AND FUTURE SCOPE

Spam filters detect unsolicited, unwanted, and virus-infested email (called spam) and stop it from getting into email inboxes. In today’s world, spam filtering is a must to protect your business. Spam is not going away. It is estimated that 70 percent of all email sent globally is spam, and the volume of spam continues to grow.

In the proposed email system, spam filters are applied to both inbound email (email entering the network) and outbound email (email leaving the network).

Spam filters use “heuristics” methods, which means that each email message is subjected to a set of predefined rules (algorithms). Each rule assigns a numerical score to the probability of the message being spam, and if the score passes a certain threshold the email is flagged as spam and blocked from going further. There are different types of spam filters for different criteria:

- Content filters – parse the content of messages, scanning for words that are commonly used in spam emails.
- Header filters – examine the email header source to look for suspicious information (such as spammer email addresses).
- Blocklist filters – stop emails that come from a blocklist of suspicious IP addresses. Some filters go further and check the IP reputation of the IP address.
- Rules-based filters – apply customized rules designed by the organization to exclude emails from specific senders, or emails containing specific words in their subject line or body.

Any spam filtering solution cannot be 100 percent effective. However, a business email system without spam filtering is highly vulnerable, if not unusable. It is important to stop as much spam as you can, to protect your network from the many possible risks: viruses, phishing attacks, compromised web links and other malicious content. Spam filters also protect your servers from being overloaded with non-essential emails, and the worse problem of being infected with spam software that may turn them into spam servers themselves. By preventing spam email from reaching your employees' mailboxes, spam filters give an additional layer of protection to your users, your network, and your business.

The project can be accessed and view from my git repository at: https://github.com/tush7301/Mail_system

REFERENCES

- https://scikit-learn.org/stable/modules/naive_bayes.html
- https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering
- https://en.wikipedia.org/wiki/Naive_Bayes_classifier
- https://www.geeksforgeeks.org/naive-bayes-classifiers/
- https://towardsdatascience.com/how-to-build-and-apply-naive-bayes-classification-for-spam-filtering-2b8d3308501
- https://github.com/ShubhamPy/Spam-Classifier