**GOVERNMENT ARTS AND SCIENCE COLLEGE (CO-EDUCATION) SANKARANKOVIL, TENKASI DISTRICT**

**MACHINE LEARNING WITH PHYTHON**

**APRIL-2023**

OPTIMIZING SPAM FILTERING:

TEAM MEMBERS:

TEAM ID: NM2023TMID16083

TEAM LEADER: **J.MANI KANDAN**

# REG.NO: 20201361506113

TEAM MEMBER 1: **M.AJITH**

# REG.NO:20201361506102

TEAM MEMBER 2: **A.SUKUMAR**

# REG.NO:20201361506124

TEAM MEMBER 3: M.**MADHAVAN** REG.NO:20201361506109

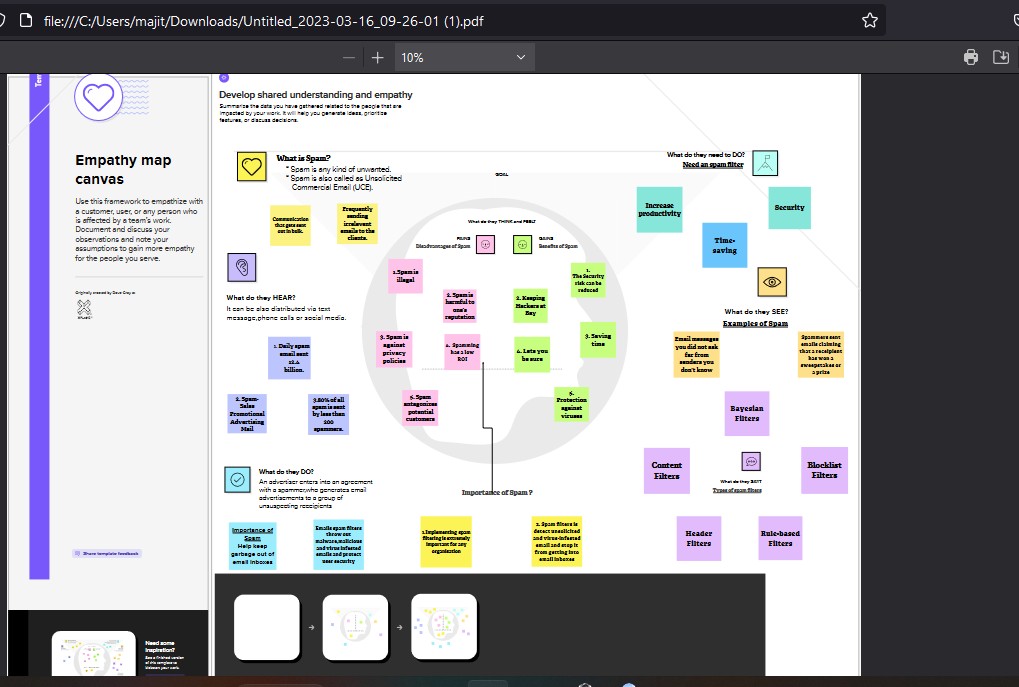
|  |  |  |
| --- | --- | --- |
| S.No | Page No. | Contents |
| 1. | 4 | INTRODUCTION |
| 2. | 6 | PROBLEM DEFINITION & DESIGN THINKING |
| 3. | 9 | RESULT |
| 4. | 13 | ADVANTAGES & DISADVANTAGES |
| 5. | 15 | APPLICATIONS |
| 6. | 21 | CONCLUSION |
| 7. | 25 | FUTURE SCOPE |
| 8. | 28 | APPENDIX |

1. **INTRODUCTION**
   1. Overview

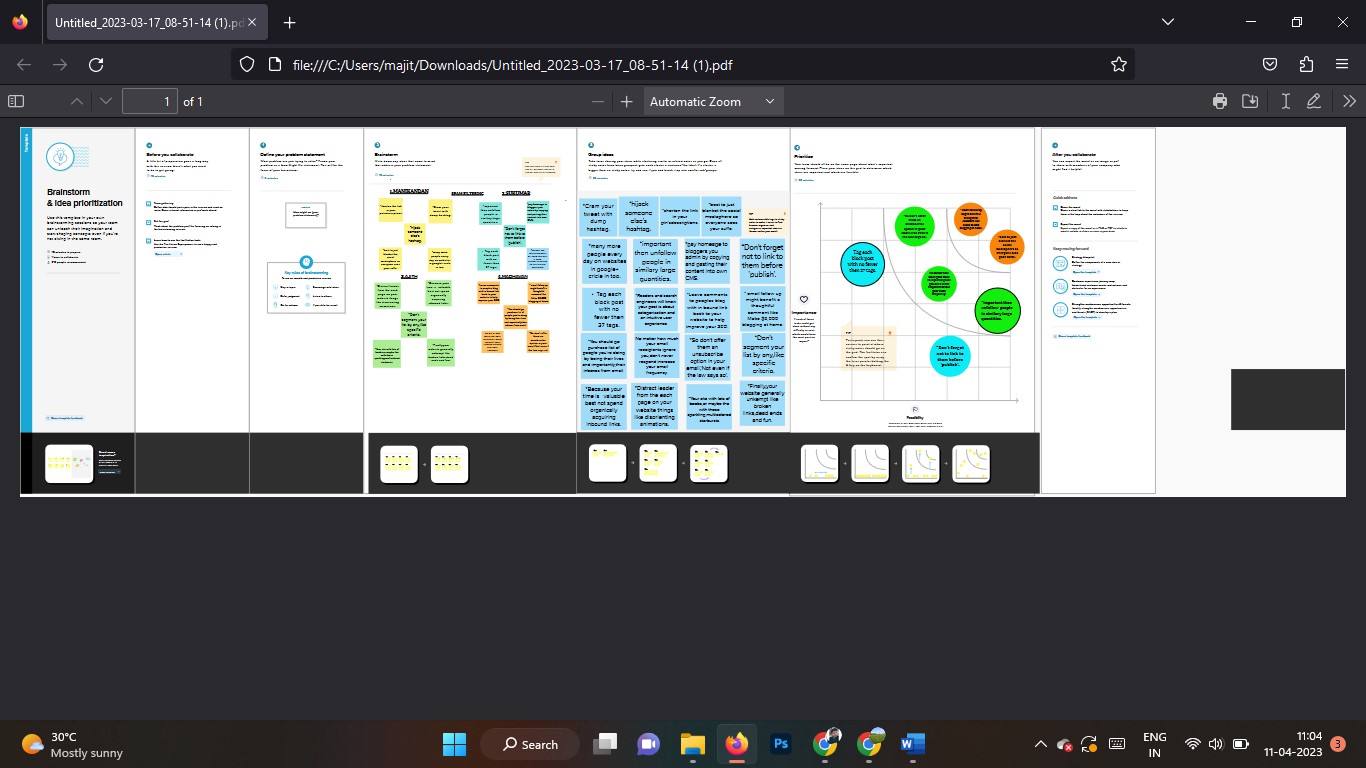
* Spam is any kind of unwanted, unsolicited digital communication that gets sent out in bulk. Often spam is sent via email, but it can also be distributed via text messages, phone calls, or social media.
* Spam is annoying, but it’s also a threat. While many of us might think we’re savvy enough to recognize any form of it, spammers regularly update their methods and messages to trick potential victims. The reality is that we’re all constantly under attack from cybercriminals and the proof is in your inbox.
* Spam is not an acronym for a computer threat, although some have been proposed (stupid pointless annoying malware, for instance). The inspiration for using the term “spam” to describe mass unwanted messages is a Monty Python skit in which the actors declare that everyone must eat the food Spam, whether they want it or not.
  1. purpose:
* The purpose of spam filtering is to identify and separate unwanted and messages from legitimate and relevant messages in electronic communica systems.such as email, text messaging, or instant messaging. Spam messages often contain fraudulent or malicious content, including phishing scams, or fraudulent offers, and can waste time, money, and resources for individuals andorganizations that receive them.
* Spam filters use various techniques, such as pattern recognition, content an sender reputation, and user feedback, to detect and block or quarantine spam messages, while allowing legitimate messages to reach their intended recipient The goal of spam filtering is to improve the quality and security of electronic communication and protect users from unwanted or harmful content.
* The spammer’s phone number is reported to Google. Google gets a copy of the last 10 incoming messages from the spammer so we can improve spam detection. In a group conversation, this can include messages from users other than the spammer.
* Spamming remains economically viable because advertisers have no operating costs beyond the management of their mailing lists, servers, infrastructures, IP ranges, and domain names, and it is difficult to hold senders accountable for their mass mailings.

***Problem Definition & Design Thinking***

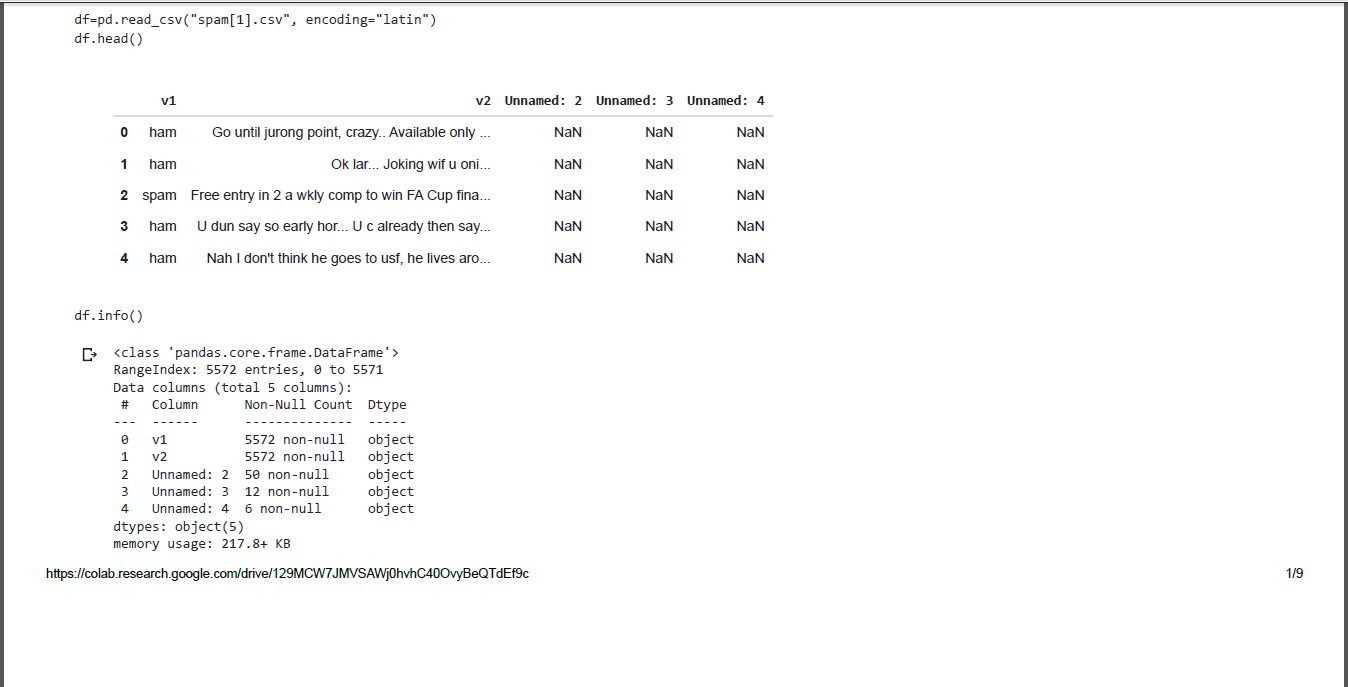
**2.1 Empathy Map**



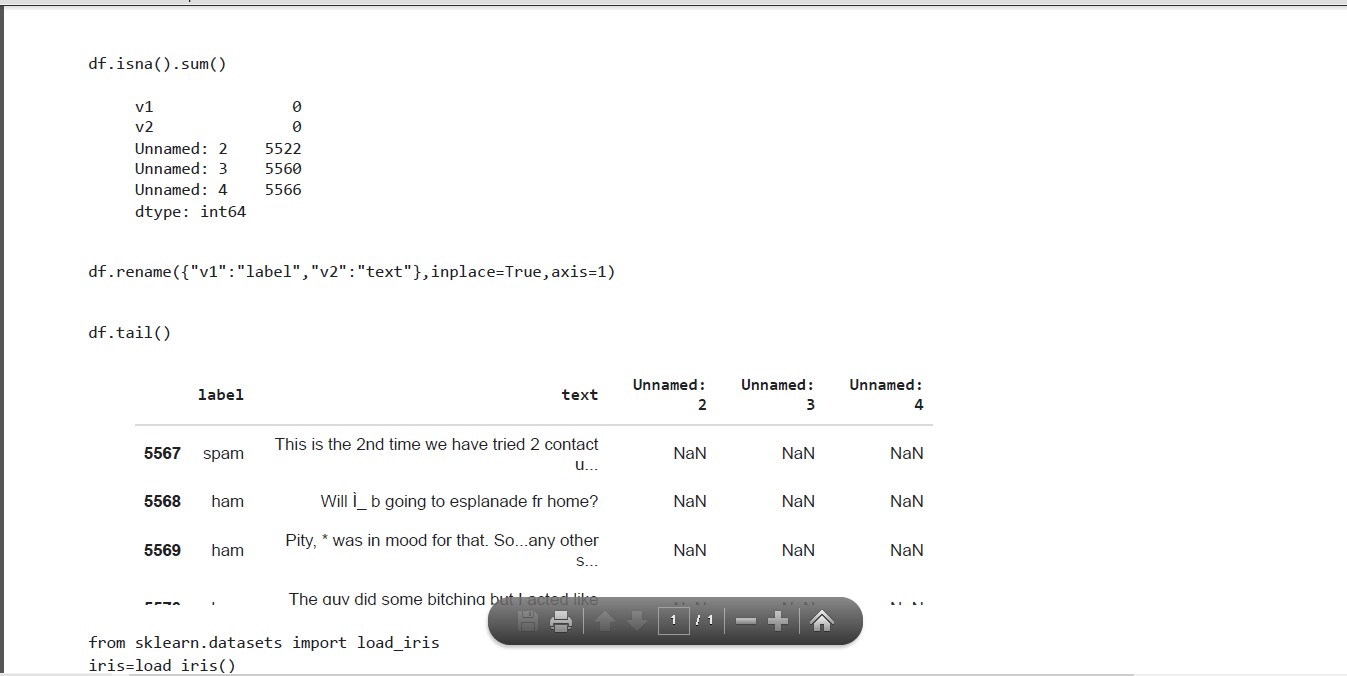
## Ideation & Brainstorming Map

******

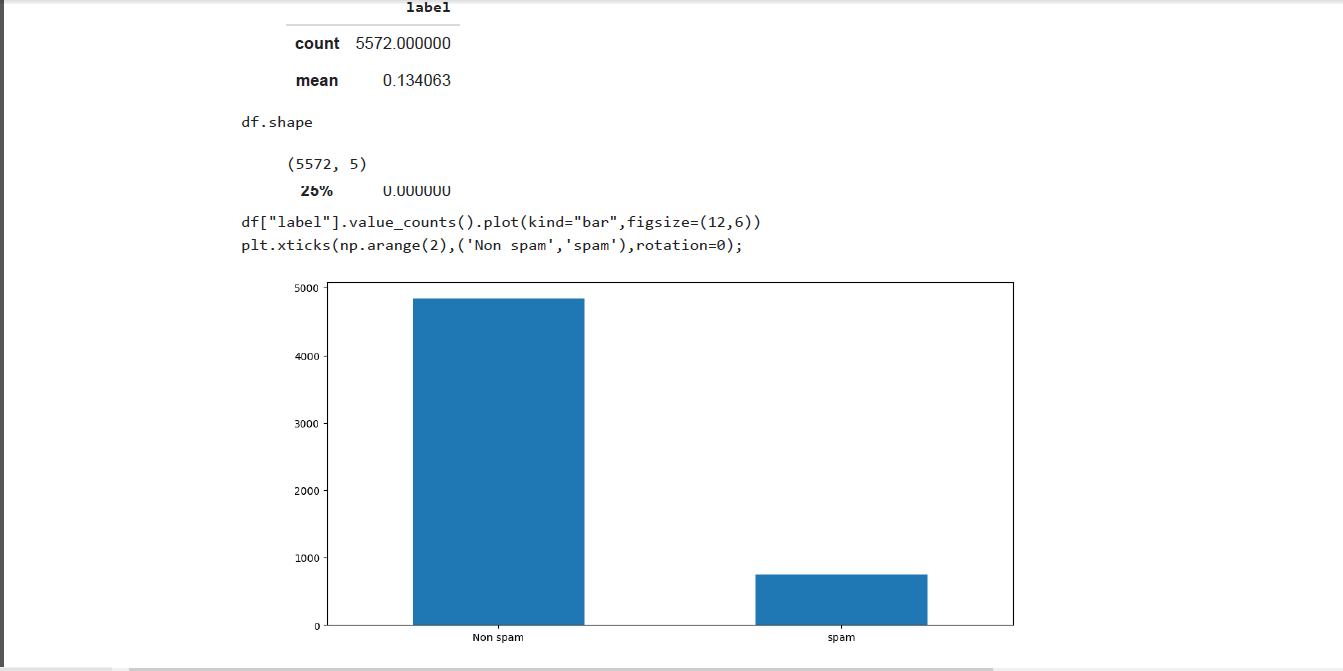
***Result:***



Handling The Missing value



Visual Analysis



## Advantages of spam

* + - Compatibility.
    - Spam Filter Relays.
    - Affordability and Pricing Flexibility.
    - Easy Maintenance.
    - Scalability.
    - Freeing Up Bandwidth.

With the benefits of email spam filters, the security risk can be reduced since the user gets in hand the emails that have gone through various spam checks. Moreover, these email spam filters throw out malware, malicious, and virus- infested emails and protect user

## Disadvantages of spam

* + - Spam antagonizes potential customers.
    - Lots of people become nervous when they receive unwanted emails considered as spam. It's due to the fact that they get hundreds of emails every day which they have to look through and they don't want to waste their time and effort on useless information coming into their inbox.
    - Spam is unethical. Spammers don’t only violate laws and people’s privacy but they also steal their money. Most email users dial into Internet providers who charge on an hourly or even minute basis. Spammers take hold of valuable Internet resources by sending junk mail but sending it “postage due”.
    - Email is considered a quick and cost-effective way of connecting with existing and potential customers.
    - Understanding the advantages and disadvantages of email marketing can help you devise a proper marketing plan that brings in more revenue.
    - While several pros and cons of email marketing exist, it all comes down to how you utilize it.
    - You need to craft an effective marketing strategy that considers your **audience’s expectations** and your **branding plan**. Keeping your message consistent and focused on the customers’ needs is critical.
    - f youí email list is not well-updated, you íun the íisk of iííitating youí customeís with iííelevant emails.



***Application***

* A spam filter is a program used to detect unsolicited, unwanted and virus - infected emails and prevent those messages from getting to a user's inbox. Like other types of filtering programs, a spam filter looks for specific criteria on which to base its judgments.
* Internet service providers free online email services and businesses use filtering tools to minimize the risk of distributing spam. For example, one of the simplest and earliest versions of spam filtering, like the one that was used by Microsoft's Hotmail, was set to watch out for particular words in the subject lines of messages. An email was excluded from the user's inbox whenever the filter recognized one of the specified words.
* Spam to a private email can cause havoc throughout the system. Nowadays, it has created many problems in business life, such as occupying network bandwidth and the space in users’ mailboxes. Research has been conducted in this area to resolve this issue and spam detection systems (SDS) have been developed to monitor spammers and filter email activities by identifying patterns in email messages, thus improving the tool to detect spam
* By finding patterns in email communications, spam detection systems (SDS) have been developed to keep track of spammers and filter email activity. SDS has also enhanced the tool for detecting spam by reducing the rate of false positives and increasing the accuracy of detection.

***Different types of spam filters***

* + - **Content filters** – Content filters scan email content for words or phrases that are commonly used in spam emails.
    - **Header filters** – Header filters scan email headers for spammer email addresses.
    - **Blocklist filters** – Blocklist filters check the IP addresses of email senders and stop emails from suspicious IP addresses.
    - **Rule-based filters** – Rule-based filters are entirely customizable and enable organizations and businesses to filter emails from specific senders or emails containing specific word phrasing.

## Applications of spam

* + - Spam filtering is a technique used to identify and remove unwanted or unsolicited email messages, commonly known as spam, from email inboxes. The primary application of spam filtering is to improve the quality of communication by preventing irrelevant or harmful messages from reaching the recipient.
    - Some common applications of spam filtering include:
    - Email security: Spam filters are used to protect email systems from spam- related security threats, including phishing attacks, malware, and viruses.
    - Productivity: Spam filtering improves productivity by reducing the amount of time spent by email users on sorting through irrelevant messages.
    - Reputation management: Spam filtering helps to maintain a positive brand image by preventing spam messages from reaching customers and business partners.
    - Regulatory compliance: Spam filtering is required by many regulatory bodies to ensure compliance with laws and regulations related to data privacy and security.
    - Cost savings: Spam filtering can save organizations money by reducing the resources required to manage and maintain email systems.
    - Overall, spam filtering is an essential tool for improving email communication and ensuring the security and privacy of digital communication.
    - Spam filtering is a widely used technology that can be applied to various fields of technology. Here are some of the applications of spam filtering in different technologies:

* + - Email: Spam filtering is extensively used in email services to detect and block unwanted or malicious emails. Email providers use various techniques like content filtering, blacklist, and Bayesian analysis to identify spam messages and prevent them from reaching the user's inbox.
    - Social Media: Spam filtering is used in social media platforms to detect and remove spam comments, posts, and messages. It helps to maintain the quality of the platform by preventing unwanted content from being shared and distributed.
    - Messaging apps: Spam filtering is used in messaging apps like WhatsApp, Telegram, and WeChat to detect and block unwanted messages. It helps to protect users from spam messages and malicious content.
    - Web Forms: Spam filtering is used in web forms to prevent automated spam submissions. It helps to ensure that the form submissions are from legitimate users and prevent the submission of unwanted or harmful content.
    - Voice over IP (VoIP): Spam filtering is used in VoIP services like Skype and Zoom to prevent unwanted calls and messages. It helps to protect users from unwanted or malicious calls and messages.

Overall, spam filtering is a versatile technology that can be applied to different fields of technology to improve communication and protect users from unwanted or harmful content

***Conclusion:***

* + - The spam Naïve Bayes method provides faster results and decent precision over all other methods from all the techniques discussed.
    - Implementing spam filtering is extremely important for any organization. Not only does spam filtering help keep garbage out of email inboxes, it helps with the quality of life of business emails because they run smoothly and are only used for their desired purpose.
    - A spam filter is a program used to detect unsolicited, unwanted and virus- infected emails and prevent those messages from getting to a user's inbox. Like other types of filtering programs, a spam filter looks for specific criteria on which to base its judgments.
    - Detection of spam is important for securing message and e-mail communication. The accurate detection of spam is a big issue, and many detection methods have been proposed by various researchers. However, these methods have a lack of capability to detect the spam accurately and efficiently. To solve this issue, we have proposed a method for spam detection using machine learning predictive models. The method is applied for the purpose of detection of spam. The experimental results obtained show that the proposed method has a high capability to detect spam. The proposed method achieved 99% accuracy which is high as compared with the other existing methods. Thus, the results suggest that the proposed method is more reliable for accurate and on-time detection of spam, and it will secure the communication systems of messages and e-mails.
    - spam filtering is not perfect, and some legitimate emails may get caught in the filter. It is important to regularly check your spam folder and mark any mistakenly filtered emails as not spam to improve the accuracy of the filtering system.
    - Overall, spam filtering is a necessary tool in today's digital age, and continued advancements in technology will help to improve its accuracy and effectiveness.
    - It is essential to stay vigilant and take precautions to protect ourselves from spam, such as not clicking on suspicious links, using strong passwords, and enabling two-factor authentication. Additionally, reporting spam to the appropriate authorities can help to prevent spammers from continuing their malicious activities.
    - Overall, while spam remains a persistent issue, continued efforts to raise awareness, implement preventative measures, and hold spammers accountable can help to minimize its impact.
    - In conclusion, spam is an unwanted and unsolicited message sent to a large number of people, typically via email or text message. Spam messages often contain misleading information, scams, and fraudulent offers, making them a significant nuisance for users.
    - Spam can be harmful, leading to identity theft, financial loss, and malware infections. It is essential to be cautious when receiving unsolicited messages and to avoid clicking on suspicious links or downloading unknown attachments.
    - Various anti-spam technologies, such as content analysis, sender reputation scoring, and machine learning algorithms, are used to filter and block unwanted messages. However, these technologies are not perfect, and some legitimate messages may get caught in the filter.

***Future Scope:***

The future scope of spam filtering projects is quite broad and promising, as new technologies and techniques emerge to better combat the problem of spam.

Some possible areas of focus for future spam filtering projects include:

1. Improving accuracy: One of the main challenges of spam filtering is ensuring that legitimate emails are not mistakenly filtered as spam. Future projects can focus on improving the accuracy of spam filtering algorithms to reduce false positives and false negatives.
2. Adapting to new threats: As spammers develop new tactics and techniques, spam filters must evolve to keep up with these threats. Future projects can focus on detecting and blocking emerging types of spam, such as image-based or social media spam.
3. Enhancing user experience: Spam filters can sometimes inconvenience users by blocking important emails or sending them to the spam folder. Future projects can focus on creating more user-friendly spam filtering systems that are less likely to disrupt legitimate communication.
4. Integrating with other systems: Spam filtering can be integrated with other security and privacy tools, such as antivirus software, firewalls, and encryption. Future projects can focus on developing more seamless integration between these systems to provide a more comprehensive approach to digital security.
5. Multilingual support: With the growth of global communication, it is important to have spam filters that can accurately identify and filter spam messages in different languages. Future projects can focus on developing multilingual spam filtering systems that can adapt to different linguistic and cultural contexts.

In summary, the future scope of spam filtering projects is vast and varied, with many opportunities for innovation and improvement. As spam continues to be a persistent problem, continued research and development in this area are critical to ensuring a safer and more efficient digital environment.

1. The next level spam filtering in used for not only messages many sources are using in technology.
2. Spam filtering project Naïve Bayes model, ANN model, Decision Tree model, Random Forest model are imported in the above project.
3. Spam filtering HTML pages are using in the project. And the Phython Code build in above project.
4. Spam filtering scope for complete automation of spam detection with maximum efficiency.
5. Spam filtering are used in mostly some advantages and some disadvantages.

***Appendix***

1. ***Source Code***

import numpy as np import pandas as pd import matplotlib.pyplot as plt import nltk from nltk.corpus import stopwords

from nltk.stem.porter import PorterStemmer import csv

from google.colab import files uploaded = files.upload()

No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

Saving spam[1] csv to spam[1] csv

df=pd.read\_csv("spam[1].csv", encoding="latin") df.head()

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **0** | h a m | G  o u n ti l j u r o n g p o i n t, c r a z y  .. A  v a il a b l e o n l y  ..  . | **2**  N  a N | **3**  N  a N | **4**  N  a N |  |
| **1** | h a m | O  k l a r.  .. J  o k i  n | N  a N | N  a N | N  a N |

g w if u o n i.

..

v1 v2 Unnamed:

**Unnamed:**

Unnamed:

* 1. spam Free entry in 2 a wkly comp to win FA Cup

fina...

NaN NaN NaN

* 1. ham U dun say so early hor... U c already then say... NaN NaN NaN

df.info()

<class 'pandas.core.frame.DataFrame'> RangeIndex: 5572 entries, 0 to 5571 Data columns (total 5 columns):

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

# Column Non-Null Count Dtype

1. v1 5572 non-null object
2. v2 5572 non-null object
3. Unnamed: 2 50 non-null object
4. Unnamed: 3 12 non-null object
5. Unnamed: 4 6 non-null object dtypes: object(5)

memory usage: 217.8+ KB

df.isna().sum()

|  |  |  |
| --- | --- | --- |
| v1 |  | 0 |
| v2 |  | 0 |
| Unnamed: | 2 | 5522 |
| Unnamed: | 3 | 5560 |
| Unnamed: | 4 | 5566 |

dtype: int64 df.rename({"v1":"label","v2":"text"},inplace=True,axis=1) df.tail()

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **labeltextUnnamed:** | **Unnam**  **ed:** | **Unnam**  **ed:** |  |
| **2** | **3** | **4** |
| **5567** spamThis is the 2nd time we have tried  2NaN | NaN | NaN |
| **5568**hamWill Ì\_ b going to esplanade fr  home?NaN | NaN | NaN |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **5569**hamPity, \* was in mood for that.  So...anyNaN | NaN | NaN |  |
|  | | | | |

The guy did some bitching but I acted

from sklearn.datasets import load\_iris iris=load\_iris() x=iris.data

y=iris.target

from sklearn.preprocessing import LabelEncoder le = LabelEncoder() df['label']=le.fit\_transform(df['label'])

from sklearn.model\_selection import train\_test\_split x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.2)

from imblearn.over\_sampling import SMOTE sm=SMOTE(random\_state=2) x\_train\_res,y\_train\_res=sm.fit\_resample(x\_train,y\_train)

nltk.download("stopwords")

contact u... other s...

[nltk\_data] Downloading package stopwords to /root/nltk\_data... [nltk\_data] Unzipping corpora/stopwords.zip.

True

import nltk

from nltk.corpus import stopwords from nltk.stem import PorterStemmer

import re corpus=[] length=len(df)

for i in range(0,length):

text=re.sub("[^a-zA-z0-9]"," ",df["text"][i]) text=text.lower() text=text.split() pe=PorterStemmer() stopword=stopwords.words("english")

text=[pe.stem(word) for word in text if not word in set(stopword)] text="

".join(text) corpus.append(text) corpus

from sklearn.feature\_extraction.text import CountVectorizer cv=CountVectorizer(max\_features=35000)

x=cv.fit\_transform(corpus).toarray()

import pickle pickle.dump(cv,open('cv1.pkl','wb'))

|  |  |
| --- | --- |
| df.desc ribe() | **label** |
| **count** | 5572.00  0000 |
| **mean** | 0.13406  3 |
| **std** | 0.34075  1 |
| **min** | 0.00000  0 |
| **25%** | 0.00000  0 |
| **50%** | 0.00000  0 |
| **75%** | 0.00000  0 |
| **max** | 1.00000  0 |

df.shape

(5572, 5)

df["label"].value\_counts().plot(kind="bar",figsize=(12,6)) plt.xticks(np.arange(2),('Non spam','spam'),rotation=0);

from sklearn.tree import DecisionTreeClassifier model=DecisionTreeClassifier() model.fit(x\_train\_res,y\_train\_res)

DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier model1=RandomForestClassifier()

model1.fit(x\_train\_res,y\_train\_res) RandomForestClassifier RandomForestClassifier()

from sklearn.naive\_bayes import MultinomialNB model=MultinomialNB() model.fit(x\_train\_res,y\_train\_res) MultinomialNB

MultinomialNB()

from tensorflow.keras.models import Sequential from tensorflow.keras.layers import Dense

model=Sequential() x\_train.shape

model.add(Dense(units=x\_train\_res.shape[1],activation="relu",kernel\_initializer="r andom\_uniform")) model.add(Dense(units=100,activation="relu",kernel\_initializer="random\_uniform"))

model.add(Dense(units=100,activation="relu",kernel\_initializer="random\_uniform")) model.add(Dense(units=1,activation="sigmoid"))

model.compile(optimizer="adam",loss="binary\_crossentropy",metrics=['accuracy'])

generator=model.fit(x\_train\_res,y\_train\_res,epochs=10,steps\_per\_epoch+len(x\_train\_ res)//64)

y\_pred=model.predict(x\_test) y\_pred

y\_pr=np.where(y\_pred>0.5,1,0) y\_test

from sklearn.metrics import confusion\_matrix,accuracy\_score cm=confusion\_matrix(y\_test,y\_pr)

score=accuracy\_score(y\_test,y\_pr) print(cm) print('Accuracy Score Is:-',score\*100)

def new\_review(new\_review):

new\_review=new\_review

new\_review=re.sub('[^a-zA-Z]','',new\_review) new\_review=new\_review.lower() new\_review=new\_review.split() ps=PorterStemmer() all\_stopwords=stopwords.words('english')

all\_stopwords.remove('not')

new\_review=[ps.stem(word) for word in new\_review if not word in set(all\_stopwords)] new\_review=''.join(new\_review)

new\_corpus=[new\_review] new\_x\_test=cv.transform(new\_corpus).toarray() print(new\_x\_test)

new\_y\_pred=loaded\_model.predict(new\_x\_test) print(new\_y\_pred) new\_x\_pred=np.where(new\_y\_pred>0.5,1,0) return new\_y\_pred new\_review=new\_review(str(input("Enter new review...")))

from sklearn.metrics import confusion\_matrix,accuracy\_score,classification\_report cm=confusion\_matrix(y\_test,y\_pred)

score=accuracy\_score(y\_test,y\_pred) print(cm) print('Accuracy Score Is Naive Bayes|:-',score\*100)

cm=confusion\_matrix(y\_test,y\_pred) score=accuracy\_score(y\_test,y\_pred) print(cm) print('Accuracy Score Is:-',score\*100) cm1=confusion\_matrix(y\_test,y\_pred1) score1=accuracy\_score(y\_test,y\_pred1) print(cm1)

print('Accuracy Score Is:-',score1\*100)

from sklearn.metrics import confusion\_matrix,accuracy\_score cm=confusion\_matrix(y\_test,y\_pr)

score=accuracy\_score(y\_test,y\_pr) print(cm) print('Accuracy Score Is:-',score\*100)

from sklearn.metrics import confusion\_matrix,accuracy\_score cm=confusion\_matrix(y\_test,y\_pr)

score=accuracy\_score(y\_test,y\_pr) print(cm) print('Accuracy Score Is:-',score\*100)

model.save('spam.h5')

from flask import Flask,render\_template,request import pickle import numpy as np import re

import nltk

from nltk.corpus import stopwords from nltk.stem import PorterStemmer from tensorflow.keras.models import load\_model

loaded\_model=load\_model('spam.h5') cv=pickle.load(open('cv1.pkl','rb')) app=Flask(\_name\_)

def home():

return render\_template('home.html')

def prediction():

return render\_template('spam.html') def predict():

if request.method=='POST': message=request.form['message'] data=message new\_review=str(data) print(new\_review)

new\_review=re.sub('[^a-zA-Z]','',new\_review) new\_review=new\_review.lower() new\_review+new\_review.split() ps=PorterStemmer() all\_stopwords=stopwords.words('english') all\_stopwords=stopwords.remove('not')

new\_review=[ps.stem(word)for word in new\_review if not word in set(all\_stopwords)] new\_review=''.join(new\_review)

new\_corpus=[new\_review] new\_x\_test=cv.transform(new\_corpus).toarray()

print(new\_x\_test)

new\_y\_pred=loaded\_model.predict(new\_x\_test) new\_x\_pred=np.where(new\_y\_pred>0.5,1,0)

print(new\_x\_pred)

if new\_review[0][0]==1:

return render\_template('result.html',prediction="Spam") else: return render\_template('result.html',prediction="Not a Spam")

if \_name\_=="\_main\_": port=int(os.environ.get('PORT',5000)) app.run(debug=False)