Project Review-1 Report

on

Spam mail prediction using machine learning

Submitted in partial fulfillment of the requirement for the award of the degree of

B.Tech Computer Science

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Mr. R Vijay

Vibhor Singh Bisht
19SCSE1010421/19021011601

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
November, 2021

Abstract

We're all aware of the problem created by spam mails . Spam mails are a hassle.
If spam reaches a user's inbox, a recipient has to delete it manually. A person who reads 10-20 emails per day may receive in the region of 160-180 spam messages along with their business correspondence. That means that they will spend 5-6 hours per month just deleting spam, to the detriment of their productive working time.

To effectively handle the threat posed by email spams, leading email providers such as Gmail, Yahoo mail and Outlook have employed the combination of different machine learning (ML) techniques such as Neural Networks in its spam filters. These ML techniques have the capacity to learn and identify spam mail and phishing messages by analyzing loads of such messages throughout a vast collection of computers.

The language used for the code of the program is python and further with the help of machine learning the code is taught to classify the emails into spam or ham(the e-mails which are not spam) with appropriate data.

The code was compiled successfully without any errors and it gave a positive output.

INTRODUCTION

Apart from the numerous benefits and conveniences people around the world can enjoy due to the Internet, there are also multiple drawbacks. Not all of them are obvious to an average user, and perhaps only professional IT workers face them from time to time. However, there is a problem almost any Internet user has encountered at least once in a lifetime. Unlike many people might think, spam is not just an annoying but harmless email message; in fact, spam can be a dangerous tool capable of harming its recipients, and should be outlawed.

Spam can cause real damage. If you wonder how a mere electronic letter can be harmful, first recall the usual contents of spam letters. Along with intrusive commercials and newsletters from electronic shops you have used only once, every email user is also at risk of receiving spam letters advertising pornography, weapons, and other questionable content. Although it might be safe in the United States of America, the European Union, and some western countries, it can be illegal in more religious countries—especially in Islamic states. A citizen of such a country who has received a spam letter with porn can be thrown in jail for nothing. A person who never looked for such content might be accused of consuming it. This is not to mention child pornography, which is also distributed and advertised through spam messages (IFR).

Having to deal with spam day by day can be stressful. Even though spam messages usually have an “unsubscribe” link, getting off a spammer’s list requires a number of actions, such as visiting the website, acknowledging unsubscription, sending confirmation letters, typing the captcha, and so on. This might be not a problem in the case of being a target of several spammers; however, usually Internet users receive dozens of spam messages daily; unsubscribing from each of them is almost impossible. Many AOL users, who are now having hard times dealing with spam, report they are already nearing the point when their mailboxes stop being useful for them because of spam (Spam Abuse).

Spam can be harmful in yet another way. Rather often, spam emails contain viruses such as trojans, worms, unblockable ads, bitcoin miners, and other malware; opening a message with such a program can instantly

damage users' computers. Spam is a tool for all kinds of frauds; for example, inheritance frauds are popular, when a user receives a personalized and seemingly credible email from a lawyer (often with a personal website and social media accounts) informing them about inheriting a sum of money, or real estate, and requiring them to send personal data for a final check. Spam often advertises low-quality, fake, or misleading products; various self-improvement and plastic surgery services and products are also distributed through spam—no need to say they deal more harm than use (FW2.com).

All these facts speak in favor of the necessity to proclaim spam illegal. A seemingly harmless electronic message can deal real damage: involve a user in a scam, contaminate their computer with malware, or even cause accusations in illegal actions such as distributing child pornography. Therefore, this problem should not be overlooked, and spam should be outlawed.

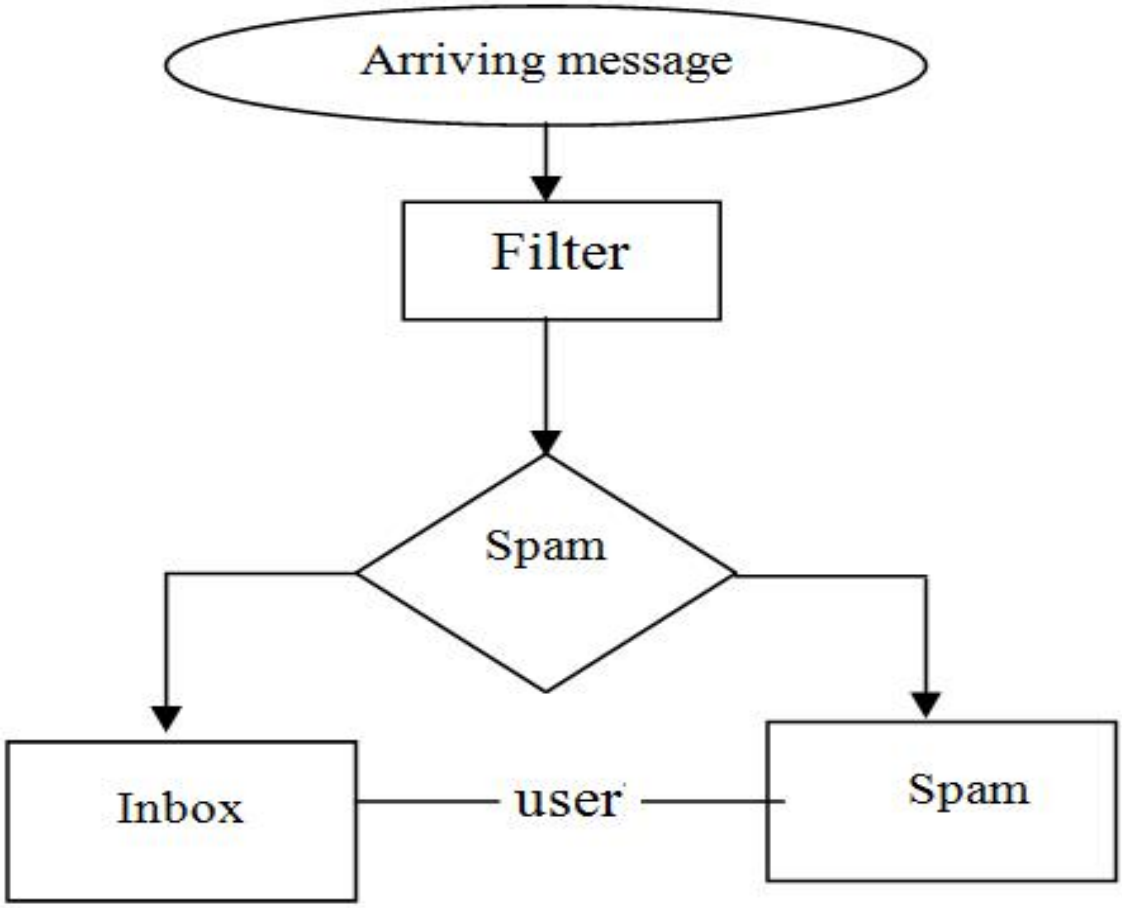
Chapter-2

Literature Survey

Spam emails are spontaneous messages to clients, which are upsetting and once in a while hurtful. There are a great deal of overview papers accessible on email spam discovery strategies. In any case, email spam recognition is similarly another region and efficient writing survey on this space is deficient. In this paper, we play out a methodical writing audit on email spam identification procedures. For that reason, we consider the accessible distributed exploration works from 2006 to 2016. We pick 17 papers for our review and checked on their pre-owned strategies, approaches and calculations, their benefits and impediments, assessment measures, conversation on datasets lastly result correlation of the examinations. Albeit, the email spam location procedures are more difficult than email spam identification strategies in view of the provincial substance, utilization of abridged words, sadly none of the current exploration tends to these difficulties. There is an enormous extent of future examination around here and this review can go about as a source of perspective point for the future course of exploration

Chapter-3

Project Design



Chapter-4

Project Implementation

```
cd
C:\Users\singh
```

In [2]:

```
import numpy as npimport pandas as pd
```

In [4]:

```
df=pd.read_csv('mail_data.csv')
```

In [5]:

```
print("The shape of data is",df.shape)df.head()
The shape of data is (5572, 2)
```

Out[5]:

	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...

In [7]:

```
from sklearn.model_selection import train_test_splitfrom
sklearn.feature_extraction.text import TfidfVectorizerfrom
sklearn.linear_model import LogisticRegressionfrom sklearn.metrics import
accuracy_score
```

In [8]:

```
#datacollection and pre-processingraw_mail_data=pd.read_csv('mail_data.csv')
```

In [9]:

```
print(raw_mail_data)
     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

[5572 rows x 2 columns]
```

In [10]:

```
#replacing null values with a null stringmail_data =
raw_mail_data.where((pd.notnull(raw_mail_data)),'')
```

In [11]:

```
mail_data.head()
```

Out[11]:

	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...

In [12]:

```
#checking the number of rows and columns in the dataframemail_data.shape
```

Out[12]:

```
(5572, 2)
```

In [13]:

```
# label spam mail as 0;  ham mail as 1;
mail_data.loc[mail_data['Category'] == 'spam', 'Category',] =
0mail_data.loc[mail_data['Category'] == 'ham', 'Category',] = 1
```

In [14]:

```
# separating the data as texts and label
X = mail_data['Message']
Y = mail_data['Category']
```

In [15]:

```
print(X)
```

```
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
```

In [16]:

```
print(Y)
```

```
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
```

In [17]:

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2,
random_state=3)
```

In [18]:

```
print(X.shape)print(X_train.shape)print(X_test.shape)
```

```
(5572,)
(4457,)
(1115,)
```

In [19]:

```
# transform the text data to feature vectors that can 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 and Y_test values as integers
Y_train = Y_train.astype('int')Y_test = Y_test.astype('int')
```

In [20]:

```
print(X_train)
```

```
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
```

In [21]:

```
print(X_train_features)
```

```
  (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
```

In [22]:

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

Out[23]:

```
LogisticRegression()
```

CHAPTER 5

Testing

In [24]:

```
# 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)
```

In [25]:

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

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

In [26]:

```
# 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)
```

In [27]:

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

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

In [28]:

```
input_mail = ["I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times"]
# convert text to feature vectorsinput_data_features = feature_extraction.transform(input_mail)
# making prediction
prediction = model.predict(input_data_features)print(prediction)


if (prediction[0]==1):
  print('Ham mail')
else:
  print('Spam mail')
[1]
Ham mail
```

Chapter-6

Conclusion and Future Work

In this review, we investigated the overall applications of spam recognizing utilizing NLP. We likewise surveyed the progression bystep interaction of the calculation and how it characterizes the mail into spam and Ham. The dataset we utilized in this paper was openly accessible, and execution measurements was too embedded to actually take a look at the model's precision. Later on, we can utilize neural organization and profound learning models to anticipate a given message is spam or not. Profound learning works very well for normal language handling; in any case, it requires a immense measure of information to give exact outcomes and to beat other customary AI calculations. Since Natural Language Processing is a generally immature region for research, further improvements can be made to the proposed framework for spam recognition and email sifting in the field of online security