EMAIL SPAM CLASSIFIER PROJECT

Submitted by:

DEBANTI ROY

ACKNOWLEDGMENT

I, Debanti Roy, would like to convey my sincere gratitude to DataTrained Academy and Flip Robo Technologies for giving me this opportunity to do this project. I would like to thank all mentors and SME's for extending their support all through the process which helped me complete this project.

E-source:

https://www.cisco.com/c/en_in/products/security/email-security/what-is-spam.html#~types-of-spam

INTRODUCTION

• Business Problem Framing

Technology has become a vital part of life in today's time. With each passing day, the use of internet increases exponentially, and with it the use of the email for the purpose of exchanging information and communicating has also increased. While e-mails are necessary for everyone, they also come with unnecessary, undesirable bulk mails, which are also called Spam Mails. Anyone with access to the internet can receive spam on their devices. Most spam emails divert people's attention away from 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 method.

• Conceptual Background of the Domain Problem

Spam email is unsolicited and unwanted junk email sent out in bulk to an indiscriminate recipient list. Typically, spam is sent for commercial purposes. It can be sent in massive volume by botnets, networks of infected computers.

Often, spam email is sent for commercial purposes. While some people view it as unethical, many businesses still use spam. The cost per email is incredibly low, and businesses can send out mass quantities consistently. Spam email can also be a malicious attempt to gain access to your computer.

Spam email can be dangerous. It can include malicious links that can infect your computer with malware (see What is malware?). Do not click links in spam. Dangerous spam emails often sound urgent, so you feel the need to act. Keep reading to learn about some of the basic spam types. Common types of spam are commercial advertisements, antivirus warnings, email spoofing, sweepstakes winners, money scams, etc. Gmail also automatically identifies spam and other suspicious emails and sends them to Spam.

• Review of Literature

Comment: A comment is something that one says which expresses his/her opinion of something or which gives an explanation of it. In the online business market, comments are an essential way to assess the quality for a product/service that are posted by an existing user.

Spam: This can be defined as unsolicited usually commercial messages (such as emails, text messages, or Internet postings) sent to a large number of recipients or posted in a large number of places.

- Motivation for the Problem Undertaken

A Spam Detector, as the name suggests, is used to detect unwanted, malicious and virus infected texts and helps to separate them from the nonspam texts. It uses a binary type of classification containing the labels such as 'ham' (nonspam) and spam. Application like this can be seen in Google Mail (GMAIL) where it segregates the spam emails in order to prevent them from getting into the user's inbox.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

- Data Sources and their formats

  What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

- Data Preprocessing Done
- **Importing the necessary libraries and packages**

First we have imported the necessary libraries.

```
#Importing necessary libraries and packages
import pandas as pd
import numpy as np

import seaborn as sns                # For Visualization
import matplotlib.pyplot as plt      # ploting package
%matplotlib inline
import matplotlib.ticker as plticker


import warnings                      # Filtering warnings
warnings.filterwarnings('ignore')
```

Then we have imported our dataset which was in CSV format and printed the shape of the dataset, i.e., the total rows and columns.

```
df=pd.read_excel(r"C:\Users\HP\Documents\Internship assignments\spam.xlsx")

print("Shape of the dataset:", df.shape)
```

```
Shape of the dataset: (5572, 5)
```

We can see the dataset has 5572 rows and 5 columns.

- **EDA**

As a part of the Exploratory Data Analysis or EDA we have printed 5 rows as sample to get a first view of our dataset.

```
df.sample(5)
```

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
2316	ham	That's cause your old. I live to be high.	NaN	NaN	NaN
1525	ham	Pls pls find out from aunt nike.	NaN	NaN	NaN
780	ham	Your opinion about me? 1. Over 2. Jada 3. Kusr...	NaN	NaN	NaN
1930	ham	Carry on not disturbing both of you	NaN	NaN	NaN
4022	ham	Well. Balls. Time to make calls	NaN	NaN	NaN

Here we can see there are 5 columns, where v1 indicates ‘spam’ or ‘ham’ referring to v2 which is the actual text or comments collected from various platforms. The last 3 columns are mostly NaN or Null values but we will find more about them in the following steps.

The next step inevitably is to find more details about our dataset. This can be easily prompted by the ‘.info()’ method.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   v1          5572 non-null   object
 1   v2          5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: object(5)
memory usage: 217.8+ KB
```

Here we can see that the last 3 columns, viz., 'Unnamed: 2', 'Unnamed:3' and 'Unnamed 4' are mostly null with each having 50, 12 and 6 non-null values only. So, it seems that these fields are unimportant and can be dropped to clean data.

```
#Data cleaning
# dropping last 3 columns as they are mostly null and are of no importance
df.drop(columns=['Unnamed: 2','Unnamed: 3','Unnamed: 4'],inplace=True)
```

With '.drop()' method we have deleted the unwanted columns.

```
# renaming the columns with meaningful names
df.rename(columns={'v1':'Category','v2':'Messages'},inplace=True)
df.sample(5)
```

	Category	Messages
3310	ham	HI DARLIN HOW WAS WORK DID U GET INTO TROUBLE?...
2634	ham	Sorry da thangam, very very sorry i am held up...
314	ham	You made my day. Do have a great day too.
243	ham	Although i told u dat i'm into baig face watch...
418	ham	Alright, I'll head out in a few minutes, text ...

We have renamed the important columns with more meaningful names.

As there were no null values we checked the data for duplicate values.

```
#checking for duplicates
df.duplicated().sum()
```

```
403
```

```
df = df.drop_duplicates(keep='first')
```

```
#checking duplicates after removal
df.duplicated().sum()
```

```
0
```

All duplicates have been removed

We found 403 duplicate entries which were removed using ".drop_duplicates(keep='first')" method.

We used the ".describe()" method to get the statistical information of the cleaned data.

```
df.describe()
```

	Category	Messages
count	5169	5169
unique	2	5169
top	ham	Go until jurong point, crazy.. Available only ...
freq	4516	1

We can see the cleaned data contains 5169 rows and most messages are 'ham' in nature counting to 4516.

When grouped separately we see the following:

```
df.groupby('Category').describe().T
```

	Category	ham	spam
Messages	count	4516	653
	unique	4516	653
	top	Go until jurong point, crazy.. Available only ...	Free entry in 2 a wkly comp to win FA Cup fina...
	freq	1	1

Observation:

-4516 ham messages.

-653 span messages.

We also checked the length of each text and found out the comment that has the maximum length.

```
df['Length'] = df['Messages'].str.len()
df.head()
```

	Category	Messages	Length
0	ham	Go until jurong point, crazy.. Available only ...	111.0
1	ham	Ok lar... Joking wif u oni...	29.0
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155.0
3	ham	U dun say so early hor... U c already then say...	49.0
4	ham	Nah I don't think he goes to usf, he lives aro...	61.0

```
df['Length'].describe()
```

```
count    5168.000000
mean       78.994969
std        58.235445
min         2.000000
25%        36.000000
50%        60.000000
75%       117.000000
max       910.000000
Name: Length, dtype: float64
```

```
df[df['Length'] == 910]['Messages'].iloc[0]
```

```
"For me the love should start with attraction.i should feel that I need her every time around me.she should be the first thing
which comes in my thoughts.I would start the day and end it with her.she should be there every time I dream.love will be then w
hen my every breath has her name.my life should happen around her.my life will be named to her.I would cry for her.will give al
l my happiness and take all her sorrows.I will be ready to fight with anyone for her.I will be in love when I will be doing the
craziest things for her.love will be when I don't have to proove anyone that my girl is the most beautiful lady on the whole pl
anet.I will always be singing praises for her.love will be when I start up making chicken curry and end up makiing sambar.life
will be the most beautiful then.will get every morning and thank god for the day because she is with me.I would like to say a l
ot..will tell later.."
```

It seems like some romeo is spamming the mail boxes.

Later we also found out the top 5 messages.

- Hardware and Software Requirements and Tools Used

Hardware required:

a. Processor: core i5 or above
b. RAM: 8 GB or above
c. ROM/SSD: 250 GB or above

Software required:

d. Anaconda 3- language used Python 3
e. Microsoft Excel Libraries: The important libraries that I have used for this project are below:

import numpy as np

It is defined as a Python package used for performing various numerical computations and processing of the multidimensional and single dimensional array elements. The calculations using Numpy arrays are faster than the normal Python array.

import pandas as pd

Pandas is a Python library that is used for faster data analysis, data cleaning and data pre-processing. The data-frame term is coming from Pandas only.

import matplotlib.pyplot as plt and import seaborn as sns

Matplotlib and Seaborn acts as the backbone of data visualization through Python.

Matplotlib: It is a Python library used for plotting graphs with the help of other libraries like Numpy and Pandas. It is a powerful tool for visualizing data in Python. It is used for creating statical interferences and plotting 2D graphs of arrays.

Seaborn: It is also a Python library used for plotting graphs with the help of Matplotlib, Pandas, and Numpy. It is built on the roof of Matplotlib and is considered as a superset of the Matplotlib library. It helps in visualizing univariate and bivariate data.

from sklearn.preprocessing import LabelEncoder

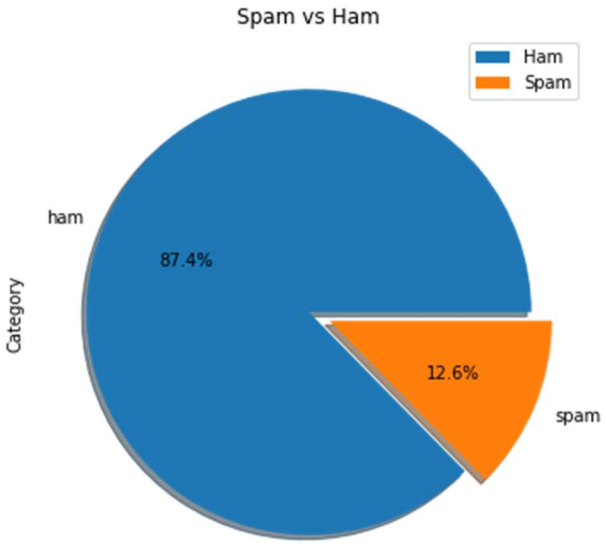
There are several encoding techniques like Label Encoder, OneHotEncoder, Ordinal Encoder.

In this project I have used OneHotEncoder technique to convert categorical data or object type data into numerical data.

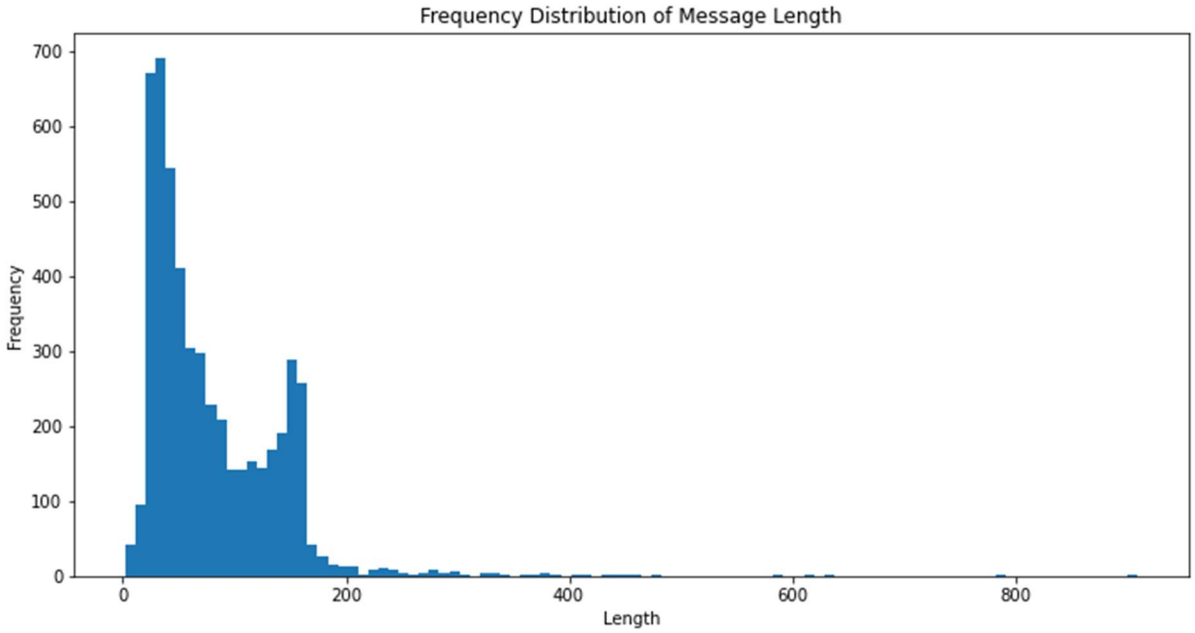
Model/s Development and Evaluation

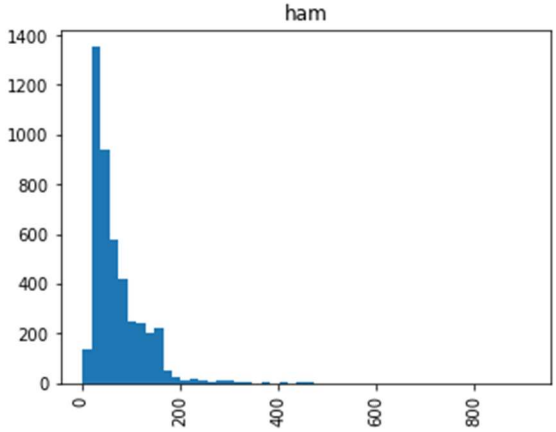
- Identification of possible problem-solving approaches (methods)
  - ❖ I have used “.drop()” function to drop unwanted entries in the columns.
  - ❖ We have renamed the important columns with more meaningful names.
  - ❖
  - ❖ Used “.drop_duplicates(keep=’first’)” method to remove duplicate entries.
  - ❖ Used “.info()” to get a detailed understanding of the data types and check for null vales.
  - ❖ To check null values I have used “.isnull().sum()” method.
  - ❖ Described the statistical details of the features using “.describe()” method.
  - ❖ Performed bivariate analysis using seaborn and matplotlib.

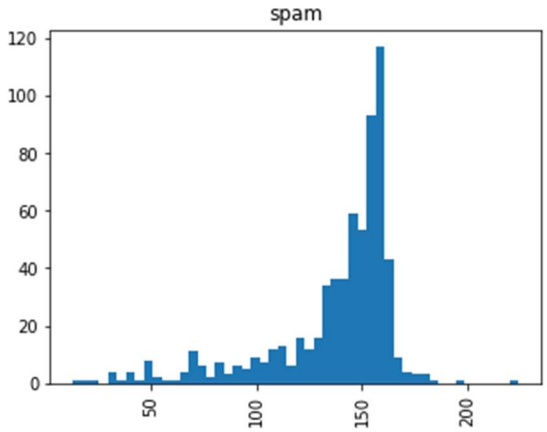
- Visualizations



We can see that 87.4% data is 'ham', that is marked in blue, while 12.6% data is 'spam', that is marked in orange.







Observation:

- The first one is a frequency distribution of the message length. Most of the message length is less than 200. Note that x-axis goes all the way to 1000ish, this must mean that there is some really long message!
- The second one is a frequency distribution in each category. Looks like spam messages are usually longer.