National University of Computer & Emerging Sciences Karachi Campus

AI (AI2002) Project Report

Classifying Emails between Normal and Spam using Naive Bayers

Section **:7A**

Group Members:

Syed Ahmed Ali (19k-1353)
Muhammad Aashir (19k-0314)
Haris Altaf (19k-1372)

1. Project Motivation

- Problem: Enron was one of the biggest energy companies at the time in the US, but it collapsed in 2000 because of a significant allegation of fraud using emails. Our task would be to find out anomalies in the emails.

- Strategic Goal: To differentiate between spam/ abnormal and normal emails in order to prevent frauds.

2. Problem Definition

- Relevant factors: words, probabilistic values.

3. Relevant Method/Model

- Output: Emails to be classified among normal and abnormal from a given data set.

- Input: Enron’s dataset of the emails. https://www.kaggle.com/code/stfkolev/gaussian-mixture-model/data?select=emails.csv

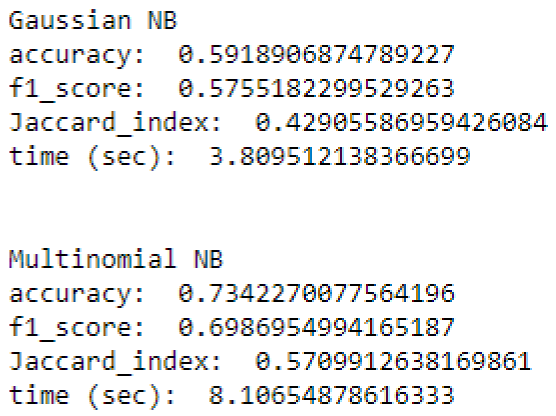
4. Performance Measurement

- Measurement of accuracy: After training the model on the data set, we will give the model a few emails to classify which we will know that they are normal or spam. And then the results will be compared.

- Minimum level of accuracy: 59%
- Maximum level of accuracy: 73%

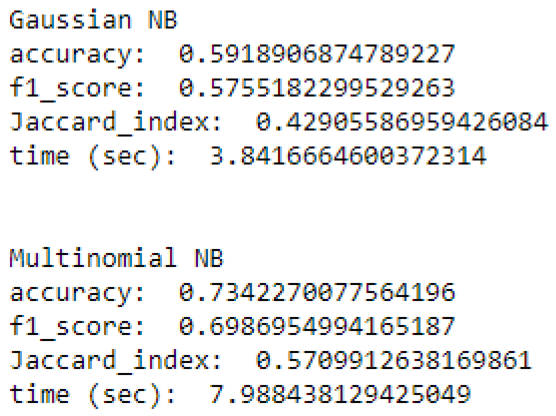
5. Risks and Dependencies
    a. Constraints
    - Hardware dependencies for training model. Have a mid range GPU and CPU.
    - Data dependencies for training model.

    b. Risks
    - Hardware constraints might not be able to take on the huge data set.
        - To tackle the situation, data set size may be lowered to a size that will not hurdle the hardware and the size will be large enough to have 90% + accuracy.
6. Run Performance Checks
    a. Classification Accuracy
        i. At the first test, the accuracy goes to 73% for Multinomial Naive Bayes.



```
Gaussian NB
accuracy:  0.5918906874789227
f1_score:  0.5755182299529263
Jaccard_index:  0.42905586959426084
time (sec):  3.809512138366699


Multinomial NB
accuracy:  0.7342270077564196
f1_score:  0.6986954994165187
Jaccard_index:  0.5709912638169861
time (sec):  8.10654878616333
```

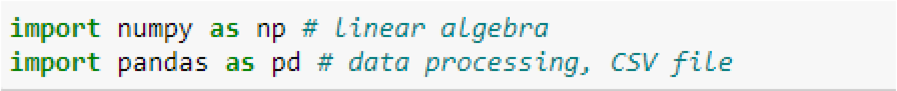
ii. At the second test, the accuracy goes to 73% for Multinomial Naive Bayes.



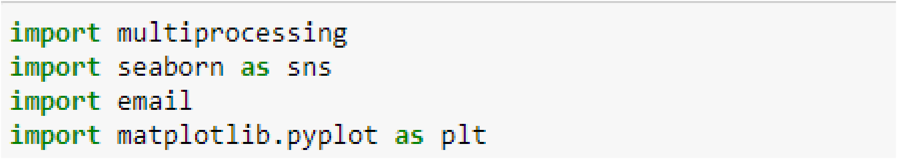
```
Gaussian NB
accuracy:  0.5918906874789227
f1_score:  0.5755182299529263
Jaccard_index:  0.42905586959426084
time (sec):  3.8416664600372314


Multinomial NB
accuracy:  0.7342270077564196
f1_score:  0.6986954994165187
Jaccard_index:  0.5709912638169861
time (sec):  7.988438129425049
```

7. Technologies
   a. Numpy
   b. Panda
   c. SKLearn
   d. Matplotlib.pyplot

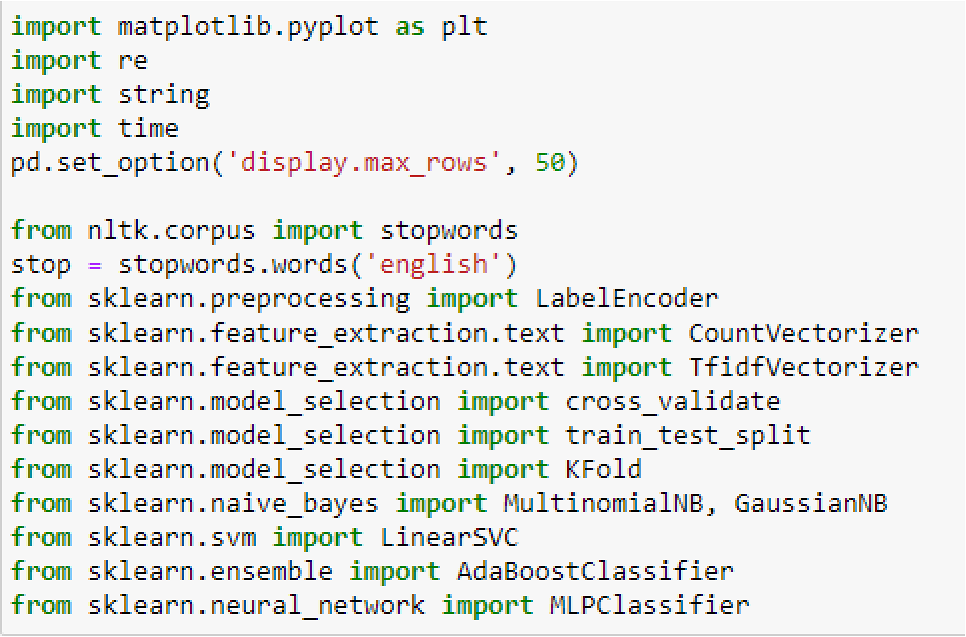
e. NLTK



```python
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file
```



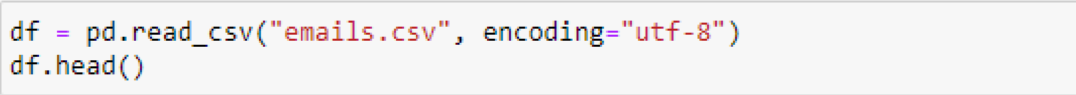
```python
import multiprocessing
import seaborn as sns
import email
import matplotlib.pyplot as plt
```



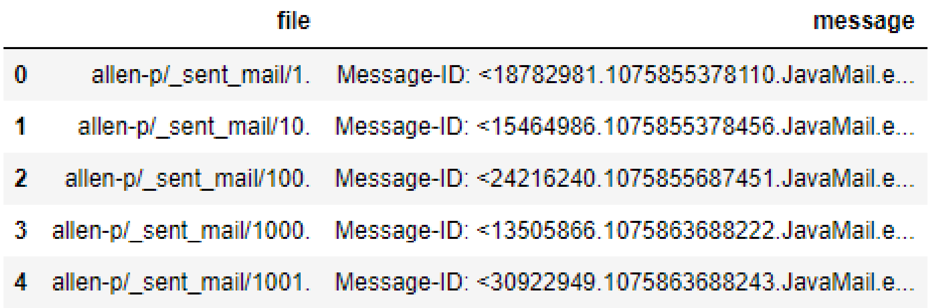
```python
import matplotlib.pyplot as plt
import re
import string
import time
pd.set_option('display.max_rows', 50)

from nltk.corpus import stopwords
stop = stopwords.words('english')
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import cross_validate
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.naive_bayes import MultinomialNB, GaussianNB
from sklearn.svm import LinearSVC
from sklearn.ensemble import AdaBoostClassifier
from sklearn.neural_network import MLPClassifier
```

8. Data Cleaning

a. Before cleaning:

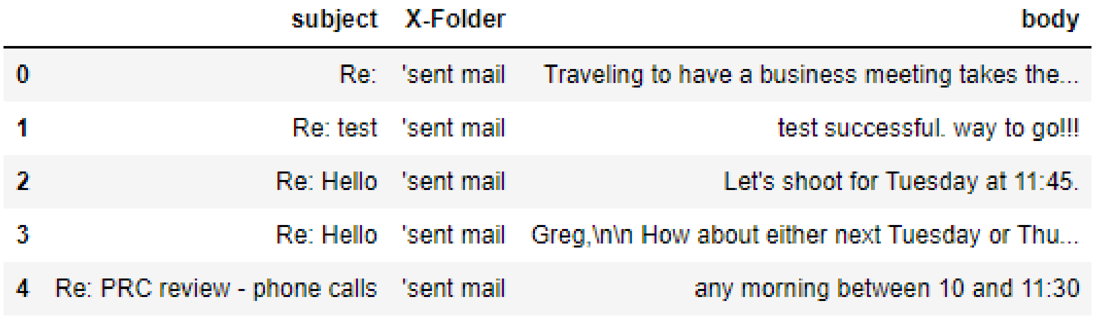


```
df = pd.read_csv("emails.csv", encoding="utf-8")
df.head()
```



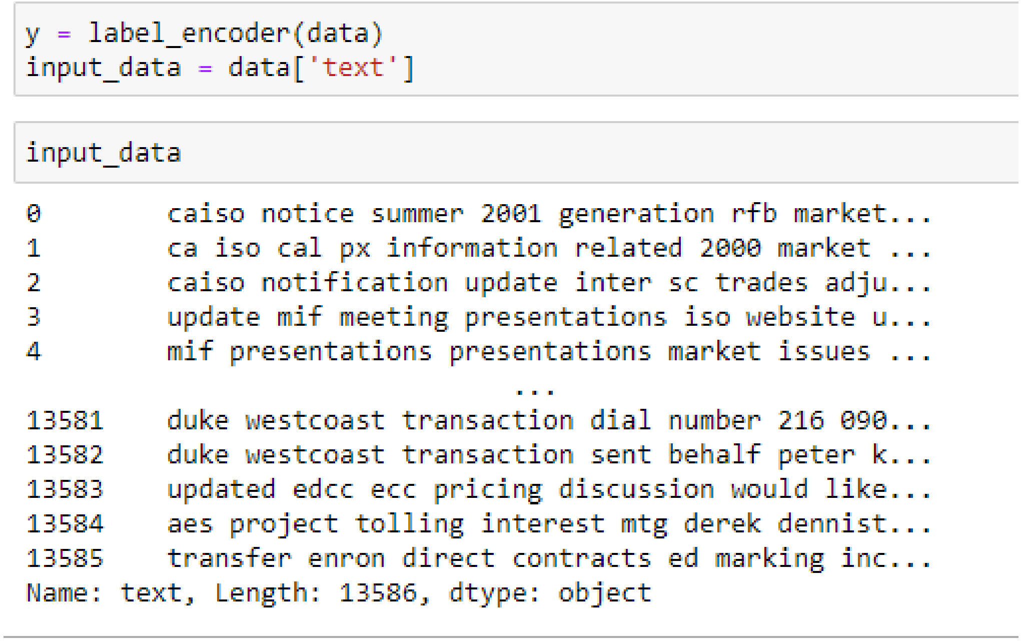
	file	message
0	allen-p/_sent_mail/1.	Message-ID: <18782981.1075855378110.JavaMail.e...
1	allen-p/_sent_mail/10.	Message-ID: <15464986.1075855378456.JavaMail.e...
2	allen-p/_sent_mail/100.	Message-ID: <24216240.1075855687451.JavaMail.e...
3	allen-p/_sent_mail/1000.	Message-ID: <13505866.1075863688222.JavaMail.e...
4	allen-p/_sent_mail/1001.	Message-ID: <30922949.1075863688243.JavaMail.e...

b. After Cleaning:



	subject	X-Folder	body
0	Re:	'sent mail	Traveling to have a business meeting takes the...
1	Re: test	'sent mail	test successful. way to go!!!
2	Re: Hello	'sent mail	Let's shoot for Tuesday at 11:45.
3	Re: Hello	'sent mail	Greg,\n\n How about either next Tuesday or Thu...
4	Re: PRC review - phone calls	'sent mail	any morning between 10 and 11:30

c. Input Data:

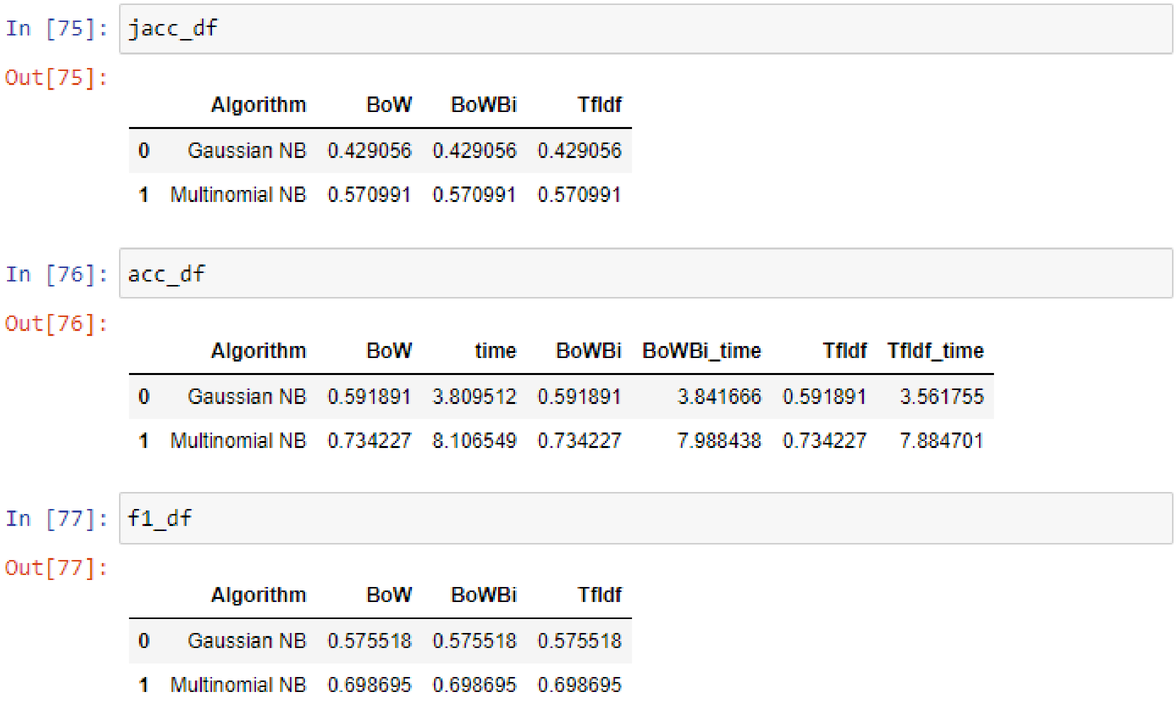


```
y = label_encoder(data)
input_data = data['text']
```

```
input_data
```

```
0        caiso notice summer 2001 generation rfb market...
1        ca iso cal px information related 2000 market ...
2        caiso notification update inter sc trades adju...
3        update mif meeting presentations iso website u...
4        mif presentations presentations market issues ...
                               ...
13581    duke westcoast transaction dial number 216 090...
13582    duke westcoast transaction sent behalf peter k...
13583    updated edcc ecc pricing discussion would like...
13584    aes project tolling interest mtg derek dennist...
13585    transfer enron direct contracts ed marking inc...
Name: text, Length: 13586, dtype: object
```

9. Results:



In [75]:

```
jacc_df
```

Out[75]:

	Algorithm	BoW	BoWBi	TfIdf
0	Gaussian NB	0.429056	0.429056	0.429056
1	Multinomial NB	0.570991	0.570991	0.570991

In [76]:

```
acc_df
```

Out[76]:

	Algorithm	BoW	time	BoWBi	BoWBi_time	TfIdf	TfIdf_time
0	Gaussian NB	0.591891	3.809512	0.591891	3.841666	0.591891	3.561755
1	Multinomial NB	0.734227	8.106549	0.734227	7.988438	0.734227	7.884701

In [77]:

```
f1_df
```

Out[77]:

	Algorithm	BoW	BoWBi	TfIdf
0	Gaussian NB	0.575518	0.575518	0.575518
1	Multinomial NB	0.698695	0.698695	0.698695