TEXT CLASSIFICATION

-BUILD A CLASSIFIER MODEL USING NAVIE BAYE'S ALGORITHM

PRESENTED BY:
YUVARAJ.S.S-203052
SREEMANN.S-203017
VASANTH.P-203302

INTRODUCTION

- A news article discusses current or recent news of either general interest or on a specific topic.
- For example ,political ,trade ,technology's etc.....
- Every news websites classifies the news article before when we go to a websites, so visitors can easily click on the type of news of interest.
- This leads to save the visitors time and might helpful for all.

DATASET-Text classification

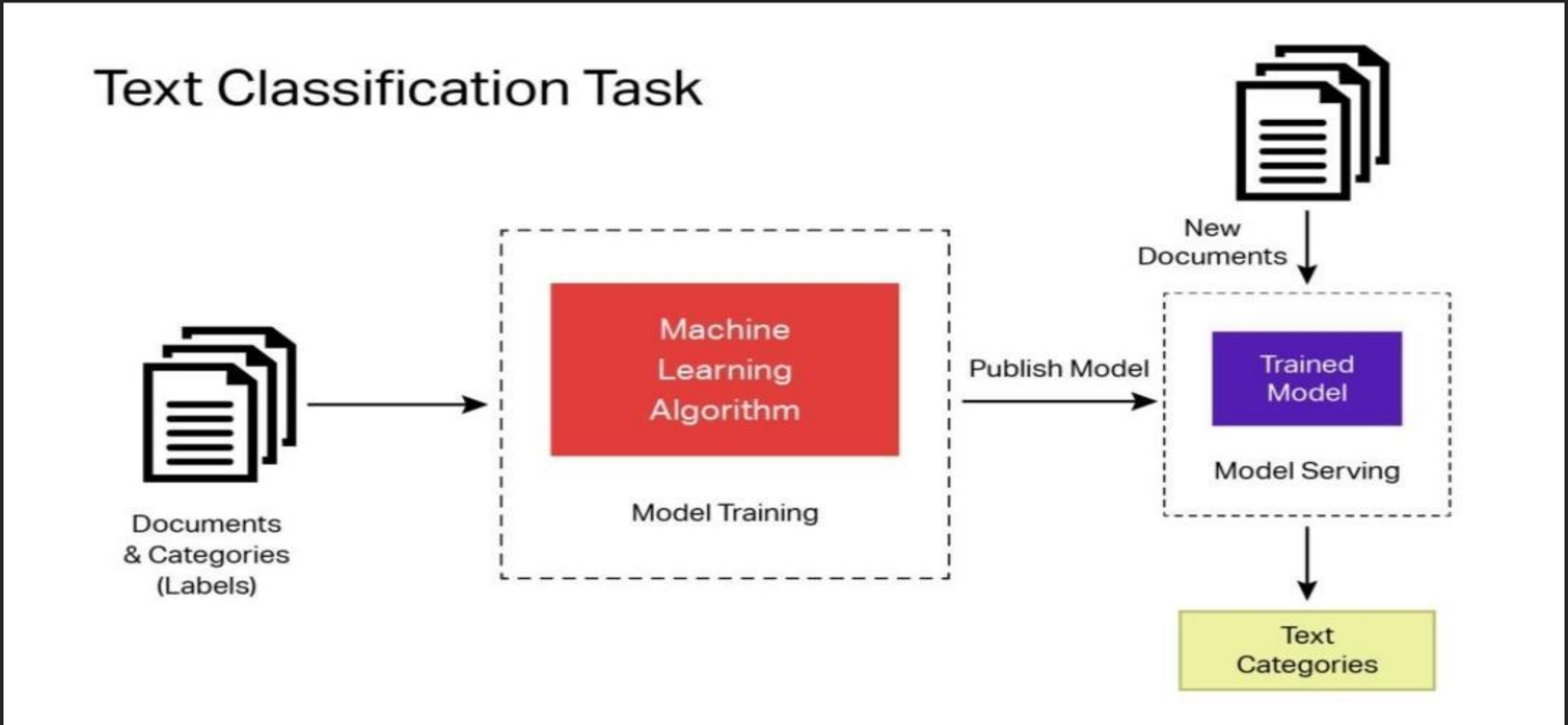
- Text classification dataset are used to categorize natural language processing texts according to the content.
- Some of the dataset providing websites for text classifications are,

1. Kaggle
2. Blogger.com
3. Newsgroup
4. Github.com

DATA CLEANINNG AND PRE PROCESSING

- Data processing is the process of transforming raw data into an understandable format.
- The quality of data should be checked before applying the machine learning or data mining algorithms.
- It is an important step because, we cannot work on the raw data.

WORK FORMAT



Text Classification Task

NAVIE BAYES ALGORITHM

- Navie bayes algorithm is a collection of classification algorithm based on bayes theorem.

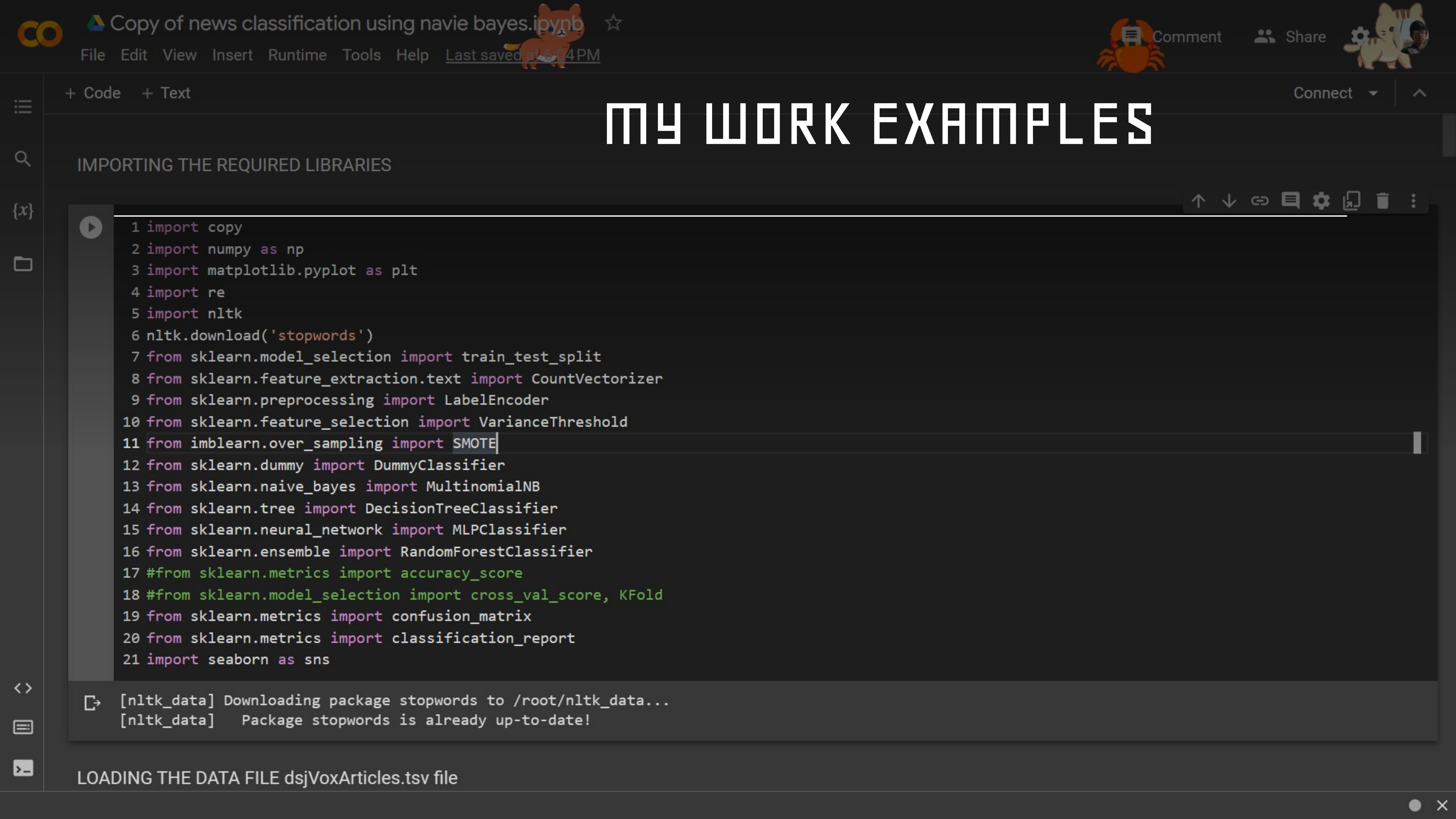
$$\boxed{P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}}$$

- A,B=events
- P[A|B]=probabilof A given B is true
- P[B|A]=probabilof B given A is true
- P[A],P[B]=the independent probabilities of A aand B

WORKING

- I can take the dataset from "Github.com" ,which is a news dataset.
- The dataset is "*dsjVoxArticles.tsv* file" and its format is TSV file.
- TSV stands for "TAB SEPERATED VALUES".
- It holds the data in the tablelled manner.

MY WORK EXAMPLES

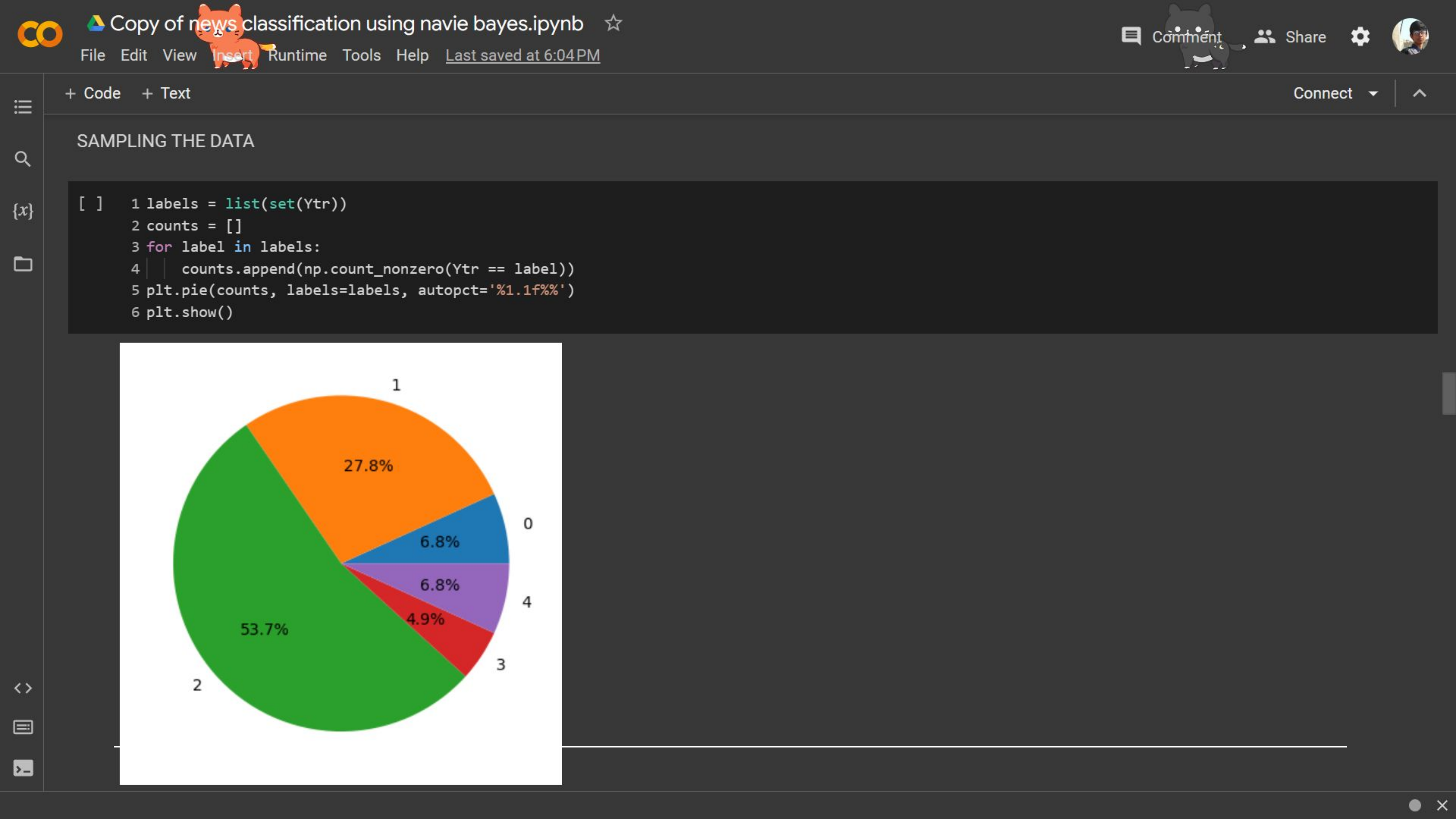


IMPORTING THE REQUIRED LIBRARIES

```python
import copy
import numpy as np
import matplotlib.pyplot as plt
import re
import nltk
nltk.download('stopwords')
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_selection import VarianceThreshold
from imblearn.over_sampling import SMOTE
from sklearn.dummy import DummyClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.ensemble import RandomForestClassifier
#from sklearn.metrics import accuracy_score
#from sklearn.model_selection import cross_val_score, KFold
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
import seaborn as sns
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

LOADING THE DATA FILE dsjVoxArticles.tsv file



SAMPLING THE DATA

```
labels = list(set(Ytr))
counts = []
for label in labels:
    counts.append(np.count_nonzero(Ytr == label))
plt.pie(counts, labels=labels, autopct='%1.1f%%')
plt.show()
```

Copy of news classification using navie bayes.ipynb

File Edit View Insert Runtime Tools Help Last saved at 6:04 PM

Comment Share

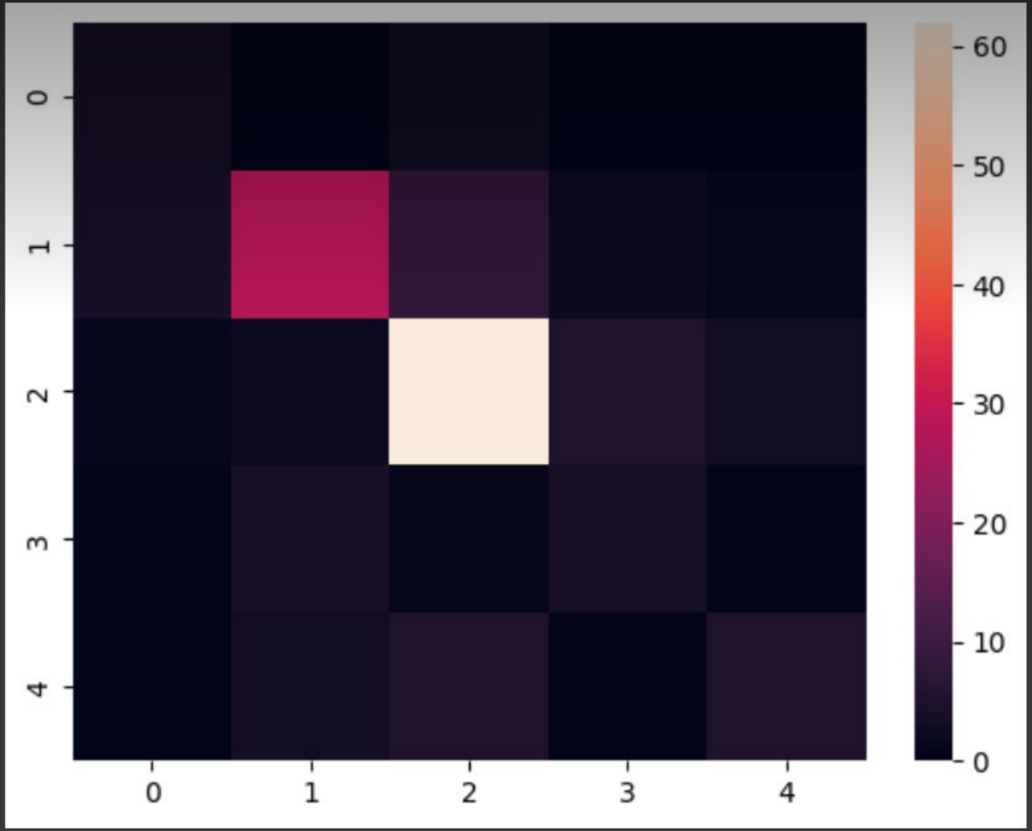
+ Code + Text Connect

FINAL

```
[ ]  Politics & Policy       0.36      0.44      0.40         9
      Science & Health       0.56      0.38      0.45        13

              accuracy                           0.71       145
             macro avg       0.58      0.58      0.57       145
          weighted avg       0.71      0.71      0.71       145

<Axes: >
```



```
[ ]  1
```

THANK YOU