importing modules

```
import pandas  as pd
import requests as req
import numpy as np
from sklearn.model_selection import train_test_split
```

urls to the datasets

```
positive_baits="https://raw.githubusercontent.com/pfrcks/clickbait-
detection/master/clickbait"
negative_baits="https://raw.githubusercontent.com/pfrcks/clickbait-
detection/master/not-clickbait"
```

A function to fetch the data from a URL

```
def my_read_function(url,label):
    response=req.get(url)
    text = response.text
    lines = text.split('\n')
    df=pd.DataFrame({'baits': lines})
    df.index.name = 'Index'
    df['Label'] = label
    #df.to_csv('baits_data.csv', index=False)
    return df
```

Read the positive and negative datasets

```
positive_dataset = my_read_function(positive_baits,'clickbait')
negative_dataset = my_read_function(negative_baits,'not-clickbait')

#reading from  the positive_dataset/click_baits_dataset
df=pd.DataFrame(positive_dataset)
df
```

```
                                                   baits      Label
Index
0      Man repairs fence to contain dog, hilarity ens...  clickbait
1      Long-Term Marijuana Use Has One Crazy Side Eff...  clickbait
2      The water from his ear trickles into the bucke...  clickbait
3      You'll Never Guess What Nick Jonas Does in the...  clickbait
4      How Cruise Liners Fill All Their Unsold Cruise...  clickbait
...                                                  ...        ...
```

```
810    OITNB's Taylor Schilling and Carrie Brownstein...  clickbait
811    Researchers have discovered the average penis ...  clickbait
812    Why it may be smart to wait to put on sunscree...  clickbait
813    What state has highest rate of rape in the cou...  clickbait
814                                                        clickbait

[815 rows x 2 columns]

#reading from  the positive_dataset/click_baits_dataset
df=pd.DataFrame(negative_dataset)
df

                                                   baits
Label
Index

0      Congress Slips CISA Into a Budget Bill That's ...  not-
clickbait
1                           DUI Arrest Sparks Controversy  not-
clickbait
2      It’s unconstitutional to ban the homeless from...  not-
clickbait
3      A Government Error Just Revealed Snowden Was t...  not-
clickbait
4      A toddler got meningitis. His anti-vac parents...  not-
clickbait
...                                                  ...            ..
.
1570   Loophole means ecstasy and loads of other drug...  not-
clickbait
1571       Astronomers Watch a Supernova and See Reruns  not-
clickbait
1572   In Indian Rapists’ Neighborhood, Smoldering An...  not-
clickbait
1573                    Strong earthquake jolts Islamabad  not-
clickbait
1574                                                       not-
clickbait

[1575 rows x 2 columns]
```

Combining the datasets

```
combined_dataset = pd.concat([positive_dataset, negative_dataset],
ignore_index=True)
df=pd.DataFrame(combined_dataset)
df.to_csv("combined.csv",index=False)
```

```
#reading the first 8 rows of the combined dataset
df.head(8)
```

```
                                               baits      Label
0  Man repairs fence to contain dog, hilarity ens...  clickbait
1  Long-Term Marijuana Use Has One Crazy Side Eff...  clickbait
2  The water from his ear trickles into the bucke...  clickbait
3  You'll Never Guess What Nick Jonas Does in the...  clickbait
4  How Cruise Liners Fill All Their Unsold Cruise...  clickbait
5                 Could Queen Elizabeth Veto Brexit?  clickbait
6      This Is the Worst Color to Paint Your Kitchen  clickbait
7                     The Shocking Truth About Sugar  clickbait
```

Shuffling the combined dataset using numpy

```
# Create an array of indices from 0 to the length of the combined
dataset
shuffled_indices = np.arange(len(combined_dataset))
# Shuffle the array of indices randomly using numpy's random.shuffle
function
np.random.shuffle(shuffled_indices)
# Use the shuffled indices to rearrange the rows of the combined
dataset, creating the shuffled dataset
shuffled_dataset = combined_dataset.iloc[shuffled_indices]
df2=pd.DataFrame(shuffled_dataset)
df2.to_csv("shuffled.csv",index=False)

#reading the first ten rows of the shuffled_dataset
df=pd.DataFrame(shuffled_dataset)
df.head(10)
```

```
                                                  baits          Label
1338  A Texas town stands divided, after armed men i...  not-clickbait
1771  Goldman Sachs banker embroiled in massive over...  not-clickbait
1307  Florida drops bill to open fracking in the Eve...  not-clickbait
2168  FIFA scandal: Sepp Blatter wins another term a...  not-clickbait
1910  Captured ISIS head of chemical weapons says th...  not-clickbait
1668  UN Removes Saudi Arabia From Human Rights Blac...  not-clickbait
138   This Behavior Is The #1 Predictor Of Divorce, ...      clickbait
2003    North Korea announces it conducted nuclear test  not-clickbait
369              BREAKING: Loch Ness Monster Found Dead      clickbait
2     The water from his ear trickles into the bucke...      clickbait
```

Split the shuffled dataset into train, validation, and test sets

```
train_data_percentage = 0.72  # 72% for training
validation_data_percentage = 0.08  # 8% for validation
test_data_percentage = 0.20  # 20% for testing

# Calculate the number of samples for each split
total_samples = len(shuffled_dataset) #number of rows in the combined
dataset
train_data_samples = int(train_data_percentage * total_samples)
#number of rows in the training dataset
validation_data_samples = int(validation_data_percentage *
total_samples) #number of rows in the validation dataset
test_data_samples = total_samples - train_data_samples -
validation_data_samples #number of rows in the testing dataset

#printing the number of rows/ samples in each dataset
print(f"total sample or the number of rows ={total_samples}
samples/rows")
print(f"validation samples or the number of rows
={validation_data_samples} samples/rows")
print(f"Training samples or the number of rows ={train_data_samples}
samples/rows")
print(f"test samples or the number of rows ={test_data_samples}
samples/rows")
```

```
total sample or the number of rows =2390 samples/rows
validation samples or the number of rows =191 samples/rows
Training samples or the number of rows =1720 samples/rows
test samples or the number of rows =479 samples/rows
```

```
# Spliting the dataset into training and the remaining data
(remaining_data)
training_data, remaining_data = train_test_split(shuffled_dataset,
test_size=(1 - train_data_percentage))

# Split the remaining data into validation and test sets
validation_data, test_data = train_test_split(remaining_data,
test_size=test_data_percentage / (test_data_percentage +
validation_data_percentage))
#saving the datasets
train_dataset=pd.DataFrame(training_data) #training dataset
train_dataset.to_csv("traning_data.csv",index=False)
validating_dataset=pd.DataFrame(validation_data) #validating  set
validating_dataset.to_csv("validating_data.csv",index=False)
testing_dataset=pd.DataFrame(test_data) #testing set
testing_dataset.to_csv('testing_dataset.csv',index=False)
```

```
#reading from each and every dataset after split
train_dataset #reading from the training dataset

                                                  baits          Label
1358  Report: Mayweather received illegal IV before ...  not-clickbait
1974  Good News. Tiger Numbers in India up From 1,40...  not-clickbait
2295          Japan plans to land rover on moon in 2018  not-clickbait
2356  Eight people injured in Canada school stabbing...  not-clickbait
186   Does the Brexit vote mean Trump will win in No...      clickbait
...                                                 ...            ...
2107       Facebook paid £4,327 corporation tax in 2014  not-clickbait
792                         Here's which team is No. 1:      clickbait
53    You may be surprised when you find out who was...      clickbait
1825  Abdullah al-Zaher: Saudi Arabia is about to be...  not-clickbait
530   Here Is the World's Ugliest Color (& It Has an...      clickbait

[1720 rows x 2 columns]

#reading from the testing dataset
testing_dataset

                                                  baits          Label
795   Jessica Alba slips on a bikini for Self's Octo...      clickbait
2134  U.S. calls Netanyahu's new media chief's remar...  not-clickbait
68    Could Lady Stoneheart finally be coming to "Ga...      clickbait
2158            Woman held for Moscow child 'beheading'  not-clickbait
2190  Scientists invent silk food wrap that's biodeg...  not-clickbait
...                                                 ...            ...
729   How to drive a six-ton potato without causing ...      clickbait
636   She Was in an Ambulance Suffering from Deadly ...      clickbait
2278  Norway's integration minister: We can't be lik...  not-clickbait
1758  Qatar National Bank allegedly hacked, data of ...  not-clickbait
2147  South Korea, U.S. to discuss missile defense; ...  not-clickbait

[479 rows x 2 columns]

#reading from the validation  dataset
validating_dataset

                                                  baits          Label
1589  Alberta passes bill banning corporate and unio...  not-clickbait
1733  Top secret "28 pages" may hold clues about Sau...  not-clickbait
1460  Obama Puts Focus on Police Success in Struggli...  not-clickbait
2205        Sweden announces first centre for raped men  not-clickbait
522   Photographs Of Cities From Space Have Shown A ...      clickbait
...                                                 ...            ...
1025  Berlin stops Airbnb renting apartments to tour...  not-clickbait
1624      China announces sanctions against North Korea  not-clickbait
1285  Google Searches for "How to Move to Canada" Sp...  not-clickbait
1621     Panama papers: China censors online discussion  not-clickbait
1475  Cologne police ordered to remove word ‘rape’ f...  not-clickbait
```

```
[191 rows x 2 columns]
```

Calculating the "target rate" for each dataset (training,validation and test)

```
train_data_target_rate = (train_dataset['Label'] ==
'clickbait').mean()
validation_data_target_rate = (validating_dataset['Label'] ==
'clickbait').mean()
test_data_target_rate = (testing_dataset['Label'] ==
'clickbait').mean()
```

what % of the three datasets is t is labeled as clickbait?

```
print(f"{train_data_target_rate.round(4)*100}% of the training data is
labeled as clickbait")
print(f"{validation_data_target_rate.round(4)*100}% of the validating
data is labeled as clickbait")
print(f"{test_data_target_rate.round(4)*100}% of the testing data is
labeled as clickbait")
```

```
35.23% of the training data is labeled as clickbait
31.94% of the validating data is labeled as clickbait
30.9% of the testing data is labeled as clickbait
```