In [1]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount
("/content/drive", force_remount=True).
```

In [2]:

```
import os
import pathlib
from pathlib import Path
os.chdir("/content/drive/My Drive/Akarshan/BERT")
!ls -l
```

```
total 21567
-rw------- 1 root root  256508 Dec 28 17:58 'Actual Compare.ipynb'
-rw------- 1 root root 8388432 Dec 26 21:48  BERT5.hdf5
-rw------- 1 root root  488019 Dec 26 22:59  Compare.ipynb
-rw------- 1 root root  476329 Dec 26 23:48 'copy EDA on results.ipynb'
-rw------- 1 root root  255088 Dec 26 23:16 'Copy of Distllbert400000.ipynb'
drwx------ 2 root root    4096 Dec  3 16:27  Data
drwx------ 4 root root    4096 Dec 28 19:29  distilbert
-rw------- 1 root root  251029 Dec 26 22:52  Distllbert400000.ipynb
-rw------- 1 root root  487917 Dec 27 00:17 'EDA on results.ipynb'
drwx------ 2 root root    4096 Dec 18 07:14 'misc model'
-rw------- 1 root root   42836 Dec 28 17:53  model.png
drwx------ 3 root root    4096 Dec  3 16:27  papers
-rw------- 1 root root   27317 Dec 29 20:42  Predict.ipynb
-rw------- 1 root root   85578 Dec 26 22:54  Roberta.ipynb
-rw------- 1 root root 5551000 Dec 26 22:48  roBERT.hdf5
-rw------- 1 root root 5551000 Dec 26 22:28  scBERT.hdf5
-rw------- 1 root root  203468 Dec 26 22:53  SciBert400k.ipynb
```

In [ ]:

```
!pip install transformers
!pip install tensorflow_addons
```

```
Collecting transformers
  Downloading transformers-4.15.0-py3-none-any.whl (3.4 MB)
     |████████████████████████████████| 3.4 MB 5.2 MB/s
Collecting sacremoses
  Downloading sacremoses-0.0.46-py3-none-any.whl (895 kB)
     |████████████████████████████████| 895 kB 56.4 MB/s
Collecting tokenizers<0.11,>=0.10.1
  Downloading tokenizers-0.10.3-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_64.whl (3.3 MB)
     |████████████████████████████████| 3.3 MB 32.3 MB/s
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.7/dist-packages (from transformers) (21.3)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-packages (from transformers) (1.19.5)
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from transformers) (4.8.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-packages (from transformers) (3.4.0)
Collecting huggingface-hub<1.0,>=0.1.0
  Downloading huggingface_hub-0.2.1-py3-none-any.whl (61 kB)
     |████████████████████████████████| 61 kB 308 kB/s
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.7/dist-packages (from transformers) (4.62.3)
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from transformers) (2.23.0)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.7/dist-packages (from transformers) (2019.12.20)
Collecting pyyaml>=5.1
  Downloading PyYAML-6.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12
```

```
_x86_64.manylinux2010_x86_64.whl (596 kB)
     |████████████████████████████████| 596 kB 55.3 MB/s
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.7/dis
t-packages (from huggingface-hub<1.0,>=0.1.0->transformers) (3.10.0.2)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/dist-
packages (from packaging>=20.0->transformers) (3.0.6)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from
importlib-metadata->transformers) (3.6.0)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-package
s (from requests->transformers) (3.0.4)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/
python3.7/dist-packages (from requests->transformers) (1.24.3)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packag
es (from requests->transformers) (2021.10.8)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (fr
om requests->transformers) (2.10)
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages (from sacr
emoses->transformers) (7.1.2)
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages (from sacrem
oses->transformers) (1.15.0)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from sac
remoses->transformers) (1.1.0)
Installing collected packages: pyyaml, tokenizers, sacremoses, huggingface-hub, transform
ers
  Attempting uninstall: pyyaml
    Found existing installation: PyYAML 3.13
    Uninstalling PyYAML-3.13:
      Successfully uninstalled PyYAML-3.13
Successfully installed huggingface-hub-0.2.1 pyyaml-6.0 sacremoses-0.0.46 tokenizers-0.10
.3 transformers-4.15.0
Collecting tensorflow_addons
  Downloading tensorflow_addons-0.15.0-cp37-cp37m-manylinux_2_12_x86_64.manylinux2010_x86
_64.whl (1.1 MB)
     |████████████████████████████████| 1.1 MB 5.1 MB/s
Requirement already satisfied: typeguard>=2.7 in /usr/local/lib/python3.7/dist-packages (
from tensorflow_addons) (2.7.1)
Installing collected packages: tensorflow-addons
Successfully installed tensorflow-addons-0.15.0
```

In [ ]:

```python
import numpy as np
import pickle
import pandas as pd
from sklearn.metrics import f1_score
from random import sample
import tensorflow as tf
from keras.models import load_model
import transformers
from transformers import pipeline,TFAutoModel, AutoTokenizer
import tensorflow_addons as tfa
import warnings
warnings.filterwarnings("ignore")
```

In [ ]:

```python
def load_data(df):
  opstr = ''

  flag = input('for using data enter 1 or for providing input enter 2: ')

  if flag=='1':
    size = int(input('sample size to use: '))

    if not(isinstance(size,int) and size>0 and size<1189321):#lenght of df
      raise Exception('sample size must belong in the range 1 to 1189321 and of type int'
)

    data=df[['SBE','Label']].sample(size,replace=False)
    # data=df[['SBE','Label']].iloc[:size]
    return data
  elif flag=='2':

    def take_input():
      value1 = str(input('input sentence below:\n'))
      value2 = int(input('input Label:'))

      try:
        if isinstance(value1,str) and value2 in [1,2]:
          pass
      except:
        raise Exception(f'{value1} has to be string and {value2} has to be 0 or 1')

      return value1,value2

    sent = []
    label = []
    flag2= True
    while(flag2):
      sentv,labelv = take_input()
      sent.append(sentv)
      label.append(labelv)
      more_input = input('Enter \'y\' for more input: ')
      if not ('y' in more_input.lower()):
        flag2 = False



    data = [[value1, value2] for value1,value2 in zip(sent,label)]

    df = pd.DataFrame(data, columns = ['SBE', 'Label'])

    return df

  else:
    print('1')
    raise Exception(f'your input {flag} is neither 1 or 2.')
```

In [ ]:

```
def load_pipe(model_name = './distilbert/model',token_name = './distilbert/tokenizer'):


  BERT = TFAutoModel.from_pretrained(model_name)

  tokenizer = AutoTokenizer.from_pretrained(token_name)

  pipe = pipeline('feature-extraction', model=BERT,
                  tokenizer=tokenizer,device=1)
  # BERT.save_pretrained('distilbert/model')
  # tokenizer.save_pretrained('distilbert/tokenizer')
  return pipe
```

In [ ]:

```
def to_predictor(data,pipe):

  if isinstance(data,pd.DataFrame):

    features = np.array(pipe(data['SBE'].to_list()),dtype='object')
    lst = []
    for idx in range(np.shape(features)[0]):
      sent_mean = np.mean(features[idx][0],axis =0)
      lst.append(sent_mean)
    feature_matrix= np.array(lst)


    test_labels = [ [0,1] if value==1 else [1,0]for value in data['Label'] ]
    test_labels = np.array(test_labels)
    return feature_matrix.astype('float32'),test_labels

  else:
    raise Exception('Data not of type pandas.DataFrame')
```

In [ ]:

```
def get_preds(feature_matrix,test_labels,model, printf1 =False,verbose=True):
  opstr = ''

  # if isinstance(feature_matrix,str):
  #   print('3')
  #   return feature_matrix

  y_pr_ts = model.predict(feature_matrix)[:,1]

  y_ts = test_labels[:,1]

  def predict_with_best_t(proba, threshould=0.718):
      predictions = []
      for i in proba:
          if i>=1-threshould:
              predictions.append(1)
          else:
              predictions.append(0)
      return predictions

  predictions = predict_with_best_t(y_pr_ts)

  if verbose:
    opstr += f'Threshold used for prediction is {1-0.718:.3f}\n'
    for t,l,p in zip(y_ts,y_pr_ts,predictions):

     opstr += f'Label: {int(t)}, Prediction:{p}, Logit: {l:.3f}\n'

  if printf1:
    f1=f1_score(y_ts,predictions )*100
    opstr += f'f1 score: {f1:.2f}\n'
  print(opstr)
```

In [ ]:

```
def model_predict(data,pipe,model):
  data = load_data(df)
  feature_matrix,test_labels = to_predictor(data,pipe)
  get_preds(feature_matrix,test_labels,model)
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

```
files = 'Data//data.csv'
df = pd.read_csv(files)
df.dropna(subset = ['SBE'],inplace =True)
```

In [ ]:

```
pipe = load_pipe()
model = load_model("BERT5.hdf5")
```

```
All model checkpoint layers were used when initializing TFDistilBertModel.

All the layers of TFDistilBertModel were initialized from the model checkpoint at ./disti
lbert/model.
If your task is similar to the task the model of the checkpoint was trained on, you can a
lready use TFDistilBertModel for predictions without further training.
```

In [ ]:

```
model_predict(df,pipe,model)#used random samples of Data
```

```
for using data enter 1 or for providing input enter 2: 1
sample size to use: 20
Threshold used for prediction is 0.282
Label: 0, Prediction:1, Logit: 0.439
Label: 0, Prediction:0, Logit: 0.279
Label: 1, Prediction:0, Logit: 0.279
Label: 0, Prediction:0, Logit: 0.279
Label: 0, Prediction:0, Logit: 0.279
Label: 0, Prediction:0, Logit: 0.279
Label: 1, Prediction:1, Logit: 0.463
Label: 1, Prediction:1, Logit: 0.463
Label: 0, Prediction:0, Logit: 0.279
Label: 0, Prediction:1, Logit: 0.318
Label: 0, Prediction:1, Logit: 0.318
Label: 0, Prediction:0, Logit: 0.279
Label: 1, Prediction:0, Logit: 0.258
Label: 0, Prediction:0, Logit: 0.279
Label: 0, Prediction:1, Logit: 0.463
Label: 0, Prediction:0, Logit: 0.039
Label: 0, Prediction:0, Logit: 0.279
Label: 0, Prediction:0, Logit: 0.279
Label: 0, Prediction:0, Logit: 0.279
Label: 0, Prediction:0, Logit: 0.279
```

In [ ]:

```
df[['SBE','Label']].sample(10)
```

Out[ ]:

| | SBE | Label |
|---|---|---|
| 66250 | This result that a contributory pension system... | 1 |
| 109640 | In these sections, we study the noncanonical H... | 0 |
| 4180 | Bernaschi et al. _CITE_ accelerated the LBM po... | 1 |
| 707069 | The resulting fields are no longer potential, ... | 1 |
| 381311 | In addition, with the new definition of the pa... | 1 |
| 770955 | If we denote by _MATH_ the smallest box (recta... | 0 |
| 82927 | From these two observations, we can calculate ... | 0 |
| 904626 | As in the _MATH_ algorithm, a node chooses the... | 0 |
| 316079 | According to the statements of the researchers... | 1 |
| 251280 | Our new approach can merge some multicast sess... | 0 |

In [ ]:

```
model_predict(df,pipe,model)
```

```
for using data enter 1 or for providing input enter 2: 2
input sentence below:
This result that a contributory pension system is friendlier to the poor than a flat-bene
```

```
fit system seems paradoxical and yet is often observed.
input Label:1
Enter 'y' for more input: y
input sentence below:
In these sections, we study the noncanonical Hamiltonian dynamics of a gyrostat in Newton
ian interaction with two spherical rigid bodies.
input Label:0
Enter 'y' for more input: y
input sentence below:
Bernaschi et al. _CITE_ accelerated the LBM portion of MURPHY, a multi-scale simulation c
ode for fluids with embedded particles that combines LBM to capture fluid flow with a mod
ified molecular dynamics solver for suspended solid particles.
input Label:1
Enter 'y' for more input: y
input sentence below:
The resulting fields are no longer potential, but remain divergence-free.
input Label:1
Enter 'y' for more input: y
input sentence below:
In addition, with the new definition of the parameters _MATH_ and _MATH_, the differentia
tion of _MATH_ can be formulated as _MATHDISP_
input Label:1
Enter 'y' for more input: y
input sentence below:
If we denote by _MATH_ the smallest box (rectangle with sides parallel to the axes) conta
ining the hole _MATH_, _MATH_, and by _MATH_ the smallest box containing _MATH_, then the
checking can be reduced to the non-absorbed holes _MATH_ for which _MATH_.
input Label:0
Enter 'y' for more input:
Threshold used for prediction is 0.282
Label: 1, Prediction:1, Logit: 0.495
Label: 0, Prediction:0, Logit: 0.279
Label: 1, Prediction:1, Logit: 0.495
Label: 1, Prediction:0, Logit: 0.258
Label: 1, Prediction:1, Logit: 0.318
Label: 0, Prediction:0, Logit: 0.279
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

```

```

In [ ]:

```

```

In [3]:

```
import pandas as pd
files = 'Data//data.csv'
df = pd.read_csv(files)
df.dropna(subset = ['SBE'],inplace =True)
```

In [20]:

```
df[['SBE','Label']].sample(10)
```

Out[20]:

| | SBE | Label |
|---|---|---|
| **990050** | In the cases of _MATH_ or _MATH_, the fitting ... | 0 |
| **1145886** | For p53, peaks in total nuclear concentration ... | 1 |
| **1094958** | This assumption can be restrictive and easily ... | 0 |
| **202128** | In this method, the right-hand side of _REF_ i... | 1 |
| **464976** | Then, for any _MATH_, it implies that _MATHDIS... | 0 |
| **484968** | So the core's radius is related to both the nu... | 0 |
| **958525** | Now consider _MATH_ and the corresponding edge... | 0 |
| **1118994** | A rigorous analytic formula is derived for the... | 1 |
| **389402** | We stress that this is only a necessary (but n... | 0 |
| **826619** | Note that, if PCA is applied, the first princi... | 0 |