```
from IPython.core.display import JSON
import os
from sklearn.model_selection import train_test_split
import tensorflow as tf
try:
  import ktrain
except:
  !pip install ktrain
  os.kill(os.getpid(), 9)
from ktrain import text
from ktrain.text import texts_from_df
import time

import pandas as pd
import numpy as np
from keras.preprocessing.text import Tokenizer


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

```
Mounted at /content/drive
```

```
gpu_info = !nvidia-smi
gpu_info = '\n'.join(gpu_info)
if gpu_info.find('failed') >= 0:
  print('Not connected to a GPU')
else:
  print(gpu_info)
```

```
Thu May 12 14:18:15 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 460.32.03    Driver Version: 460.32.03    CUDA Version: 11.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla K80           Off  | 00000000:00:04.0 Off |                    0 |
| N/A   37C    P8    25W / 149W |      0MiB / 11441MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

```
reviews = pd.read_json("/content/drive/MyDrive/Colab_Notebooks/dtsa5798/data.json")
```

```
reviews.head()
```

	category	headline	authors	
0	CRIME	There Were 2 Mass Shootings In Texas Last Week...	Melissa Jeltsen	https://w
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2...	Andy McDonald	https://
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://w
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And D...	Ron Dicker	https://w
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags...	Ron Dicker	https://w

```
reviews['combined_text'] = reviews['headline'] + ' ' + reviews['short_description']
```

```
reviews[reviews['category'].str.contains("HEALTHY LIVING")]
```

	category	headline	autl
7578	HEALTHY LIVING	To The People Who Say ‘I’m Tired’ When Someone...	The Mighty, ContributorWe disability,
7693	HEALTHY LIVING	Eating Shake Shack Made Me Feel Healthier Than...	Colleen Werner, ContributorCar Editor-at-L
7747	HEALTHY LIVING	How To Stay Updated On The News Without Losing...	Lindsay Ho
7927	HEALTHY LIVING	27 Perfect Tweets About Whole30 That Will Make...	Lindsay Ho
7934	HEALTHY LIVING	The Real Reason Your Hands Are Always Cold	Refinery29, ContributorThe #1 media bra
...	...	...	
124913	HEALTHY LIVING	Why You Need Both a 'Bouncer' and a 'Bartender...	Elizabeth Grace Saun ContributorFound
124914	HEALTHY LIVING	How Video Games Can Improve Dialogue on Mental...	Mona Shattell, Contributorr resea
124925	HEALTHY LIVING	Wake-Up Calls Inspired My Change From Overdriv...	Jane Shure, ContributorLeade Coach, Psy
124950	HEALTHY LIVING	Loving a Narcissist Without Losing Yourself	Nancy C ContributorPsychotherapist, i
124988	HEALTHY LIVING	Reasons Not to Be Happy	Mindy Utay, Contributor"Calming Cont

6694 rows × 7 columns

```
reviews['healthy'] = np.where((reviews['category'] == 'HEALTHY LIVING'), 1, 0)
```

```
sample_amount =  len(reviews[reviews["healthy"] == 1])
```

```
healthy = reviews[reviews['healthy'] == 1]
not_healthy = reviews[reviews['healthy'] == 0].sample(n=sample_amount)
```

```
review_sample = pd.concat([healthy,not_healthy])
```

```
review_sample.describe()
```

	healthy
count	13388.000000
mean	0.500000
std	0.500019
min	0.000000
25%	0.000000
50%	0.500000
75%	1.000000
max	1.000000

```
target_names = ['NOT HEALTHY LIVING','HEALTHY LIVING']
```

```
tf.keras.backend.clear_session() #to prevent training on top of training
t = text.Transformer('distilbert-base-uncased', maxlen=512, class_names=target_names)
#'roberta-base', 'distilbert-base-uncased', 'distilroberta-base', 'distilroberta-base'
```

Downloading: 100% 483/483 [00:00<00:00, 11.2kB/s]

```
train, val, preprocess = ktrain.text.texts_from_df(
    review_sample,
    "combined_text",
    label_columns=["healthy"],
    val_df=None,
    max_features=10000,
    maxlen=512,
    val_pct=0.1,
    ngram_range=0, #do you want tensforflow to only consider unigrams or combos of wo
    preprocess_mode="distilbert", #try roberta-base, bert-base-uncased, distilroberta-
    verbose=0
)
```

```
['not_healthy', 'healthy']
        not_healthy  healthy
117607          1.0      0.0
98966           0.0      1.0
112887          0.0      1.0
43076           1.0      0.0
113955          1.0      0.0
['not_healthy', 'healthy']
       not_healthy  healthy
58106          1.0      0.0
37618          1.0      0.0
50082          0.0      1.0
84081          1.0      0.0
21366          1.0      0.0
```

Downloading: 100% 232k/232k [00:00<00:00, 3.54kB/s]

Downloading: 100% 466k/466k [00:00<00:00, 1.08MB/s]

```
model = preprocess.get_classifier()
learner = ktrain.get_learner(model, train_data=train, val_data=val, batch_size=16)
# batch size is 16 or under for text, can decrease or increase to get good performance
```

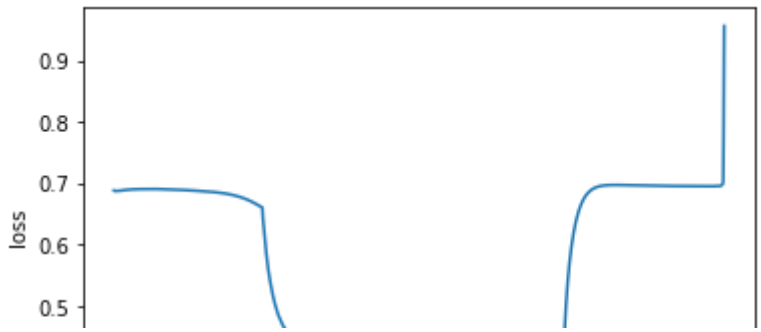
Downloading: 100% 363M/363M [00:29<00:00, 14.8MB/s]

```
learner.lr_find(max_epochs=6)
```

```
simulating training for different learning rates... this may take a few moments.
Epoch 1/6
753/753 [==============================] - 1296s 2s/step - loss: 0.6506 - accura
Epoch 2/6
753/753 [==============================] - 1278s 2s/step - loss: 0.3723 - accura
Epoch 3/6
753/753 [==============================] - 1278s 2s/step - loss: 0.4401 - accura
Epoch 4/6
753/753 [==============================] - 1268s 2s/step - loss: 0.6950 - accura
Epoch 5/6
753/753 [==============================] - 93s 121ms/step - loss: 4.0090 - accura


done.
Please invoke the Learner.lr_plot() method to visually inspect the loss plot to
```

```
learner.lr_plot()
```




```
history=learner.autofit(
    1e-4,
    checkpoint_folder='checkpoint',
    epochs=5,
    early_stopping=True
)
```

```
    olicy with max lr of 0.0001...

    s 2s/step - loss: 0.3304 - accuracy: 0.8654 - val_loss: 0.2957 - val_accuracy: 0.

     0s - loss: 0.2132 - accuracy: 0.9210Restoring model weights from the end of the
    s 2s/step - loss: 0.2132 - accuracy: 0.9210 - val_loss: 0.3191 - val_accuracy: 0.

    odel.
```

```
predictor = ktrain.get_predictor(learner.model, preproc=preprocess)
```

```
predictor.save("drive/MyDrive/MSDSTextClassification_Lab2.healthy_living")
```

```
validation = learner.validate(val_data=val, print_report=True)
```

```
              precision    recall  f1-score   support

           0       0.91      0.85      0.88       667
           1       0.86      0.91      0.88       672

    accuracy                           0.88      1339
   macro avg       0.88      0.88      0.88      1339
weighted avg       0.88      0.88      0.88      1339
```

✓ 22s completed at 11:39 AM