In [1]:
```
!pip install transformers
!pip install datasets
```

```
Collecting transformers
  Downloading transformers-4.19.1-py3-none-any.whl (4.2 MB)
     |████████████████████████████████| 4.2 MB 5.3 MB/s
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from transformers) (4.11.3)
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 36.7 MB/s
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.7/dist-packages (from transformers) (4.64.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-packages (from transformers) (3.6.0)
Collecting huggingface-hub<1.0,>=0.1.0
  Downloading huggingface_hub-0.6.0-py3-none-any.whl (84 kB)
     |████████████████████████████████| 84 kB 2.4 MB/s
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from transformers) (2.23.0)
Collecting tokenizers!=0.11.3,<0.13,>=0.11.1
  Downloading tokenizers-0.12.1-cp37-cp37m-manylinux_2_12_x86_64.manylinux2010_x86_64.whl (6.6 MB)
     |████████████████████████████████| 6.6 MB 47.8 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) (1.21.6)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.7/dist-packages (from transformers) (2019.12.20)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.7/dist-packages (from huggingface-hub<1.0,>=0.1.0->transformers)
(4.2.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from packaging>=20.0->transformers) (3.0.8)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metadata->transformers) (3.8.0)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests->transformers) (2.10)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests->transformers) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests->transformers) (2021.10.8)
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) (1.2
4.3)
Installing collected packages: pyyaml, tokenizers, huggingface-hub, transformers
  Attempting uninstall: pyyaml
    Found existing installation: PyYAML 3.13
    Uninstalling PyYAML-3.13:
      Successfully uninstalled PyYAML-3.13
Successfully installed huggingface-hub-0.6.0 pyyaml-6.0 tokenizers-0.12.1 transformers-4.19.1
Collecting datasets
  Downloading datasets-2.2.1-py3-none-any.whl (342 kB)
     |████████████████████████████████| 342 kB 5.3 MB/s
Collecting aiohttp
  Downloading aiohttp-3.8.1-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_64.whl (1.1 MB)
     |████████████████████████████████| 1.1 MB 23.9 MB/s
Requirement already satisfied: huggingface-hub<1.0.0,>=0.1.0 in /usr/local/lib/python3.7/dist-packages (from datasets) (0.6.0)
Requirement already satisfied: pandas in /usr/local/lib/python3.7/dist-packages (from datasets) (1.3.5)
Requirement already satisfied: dill in /usr/local/lib/python3.7/dist-packages (from datasets) (0.3.4)
Requirement already satisfied: packaging in /usr/local/lib/python3.7/dist-packages (from datasets) (21.3)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-packages (from datasets) (1.21.6)
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from datasets) (4.11.3)
Collecting fsspec[http]>=2021.05.0
  Downloading fsspec-2022.3.0-py3-none-any.whl (136 kB)
     |████████████████████████████████| 136 kB 74.9 MB/s
Collecting xxhash
  Downloading xxhash-3.0.0-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (212 kB)
     |████████████████████████████████| 212 kB 73.5 MB/s
Requirement already satisfied: pyarrow>=6.0.0 in /usr/local/lib/python3.7/dist-packages (from datasets) (6.0.1)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.7/dist-packages (from datasets) (4.64.0)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.7/dist-packages (from datasets) (0.70.12.2)
Collecting responses<0.19
  Downloading responses-0.18.0-py3-none-any.whl (38 kB)
Requirement already satisfied: requests>=2.19.0 in /usr/local/lib/python3.7/dist-packages (from datasets) (2.23.0)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.7/dist-packages (from huggingface-hub<1.0.0,>=0.1.0->datasets) (6.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.7/dist-packages (from huggingface-hub<1.0.0,>=0.1.0->datasets)
(4.2.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-packages (from huggingface-hub<1.0.0,>=0.1.0->datasets) (3.6.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from packaging->datasets) (3.0.8)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests>=2.19.0->datasets) (2021.10.8)
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>=2.19.0->datasets)
(1.24.3)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests>=2.19.0->datasets) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests>=2.19.0->datasets) (2.10)
Collecting urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1
  Downloading urllib3-1.25.11-py2.py3-none-any.whl (127 kB)
     |████████████████████████████████| 127 kB 74.2 MB/s
Collecting multidict<7.0,>=4.5
  Downloading multidict-6.0.2-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (94 kB)
     |████████████████████████████████| 94 kB 3.2 MB/s
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp->datasets) (21.4.0)
Collecting frozenlist>=1.1.1
  Downloading frozenlist-1.3.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (144 kB)
     |████████████████████████████████| 144 kB 75.6 MB/s
Requirement already satisfied: charset-normalizer<3.0,>=2.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp->datasets) (2.0.12)
Collecting yarl<2.0,>=1.0
  Downloading yarl-1.7.2-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_64.whl (271 kB)
     |████████████████████████████████| 271 kB 73.2 MB/s
Collecting async-timeout<5.0,>=4.0.0a3
  Downloading async_timeout-4.0.2-py3-none-any.whl (5.8 kB)
Collecting asynctest==0.13.0
  Downloading asynctest-0.13.0-py3-none-any.whl (26 kB)
Collecting aiosignal>=1.1.2
  Downloading aiosignal-1.2.0-py3-none-any.whl (8.2 kB)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metadata->datasets) (3.8.0)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from pandas->datasets) (2022.1)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from python-dateutil>=2.7.3->pandas->datasets) (1.15.0)
Installing collected packages: multidict, frozenlist, yarl, urllib3, asynctest, async-timeout, aiosignal, fsspec, aiohttp, xxhash, responses, datase
ts
  Attempting uninstall: urllib3
    Found existing installation: urllib3 1.24.3
    Uninstalling urllib3-1.24.3:
      Successfully uninstalled urllib3-1.24.3
```

```
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the follo
wing dependency conflicts.
datascience 0.10.6 requires folium==0.2.1, but you have folium 0.8.3 which is incompatible.
Successfully installed aiohttp-3.8.1 aiosignal-1.2.0 async-timeout-4.0.2 asynctest-0.13.0 datasets-2.2.1 frozenlist-1.3.0 fsspec-2022.3.0 multidict-
6.0.2 responses-0.18.0 urllib3-1.25.11 xxhash-3.0.0 yarl-1.7.2
```

In [2]:
```
import pandas as pd
import numpy as np
import torch
from torch import cuda
import random
import os
import torch
from torch import nn
from transformers import Trainer
from collections import Counter
from sklearn.model_selection import train_test_split
from sklearn.metrics import f1_score, classification_report
from sklearn.utils.class_weight import compute_class_weight
import datasets
from datasets import Dataset, load_metric
import transformers
from transformers import AutoTokenizer
from transformers import AutoModelForSequenceClassification, TrainingArguments, Trainer, DataCollatorWithPadding
from transformers import RobertaTokenizerFast, TFRobertaForSequenceClassification

print(torch.__version__)
print(transformers.__version__)
```

```
1.11.0+cu113
4.19.1
```

In [3]:
```
# To add your own Drive Run this cell.
from google.colab import drive
drive.mount('/content/gdrive')
```

```
Mounted at /content/gdrive
```

In [4]:
```
train_path = '/content/gdrive/MyDrive/Colab Notebooks/nlp-getting-started/train.csv'
train_data = pd.read_csv(train_path)
train_data = train_data[['text', 'target']]
train_data.rename(columns={"target": "label"}, inplace=True)

test_path = '/content/gdrive/MyDrive/Colab Notebooks/nlp-getting-started/train.csv'
test_data = pd.read_csv(test_path)
test_data = test_data[['text']]
test_data.rename(columns={"target": "label"}, inplace=True)
```

In [5]:
```
train_data, dev_data = train_test_split(train_data, test_size=0.1, shuffle=True, stratify=train_data['label'])

print("Train dataset labels count = ", Counter(train_data['label']))
print("Dev dataset labels count = ", Counter(dev_data['label']))
#print("Test dataset labels count = ", Counter(test_data['target'])) #test dataset does not contain the target label
```

```
Train dataset labels count =  Counter({0: 3907, 1: 2944})
Dev dataset labels count =  Counter({0: 435, 1: 327})
```

In [6]:
```
from transformers import RobertaForSequenceClassification
from transformers import RobertaTokenizerFast
model_checkpoint = 'roberta-large'
tokenizer = RobertaTokenizerFast.from_pretrained(model_checkpoint)
model = RobertaForSequenceClassification.from_pretrained(model_checkpoint)
```

```
Downloading:   0%|          | 0.00/878k [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/446k [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/1.29M [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/482 [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/1.33G [00:00<?, ?B/s]

Some weights of the model checkpoint at roberta-large were not used when initializing RobertaForSequenceClassification: ['lm_head.layer_norm.weight', 'lm
_head.dense.bias', 'roberta.pooler.dense.bias', 'lm_head.decoder.weight', 'lm_head.dense.weight', 'lm_head.layer_norm.bias', 'roberta.pooler.dense.weigh
t', 'lm_head.bias']
- This IS expected if you are initializing RobertaForSequenceClassification from the checkpoint of a model trained on another task or with another archit
ecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).
- This IS NOT expected if you are initializing RobertaForSequenceClassification from the checkpoint of a model that you expect to be exactly identical (i
nitializing a BertForSequenceClassification model from a BertForSequenceClassification model).
Some weights of RobertaForSequenceClassification were not initialized from the model checkpoint at roberta-large and are newly initialized: ['classifier.
out_proj.weight', 'classifier.dense.bias', 'classifier.dense.weight', 'classifier.out_proj.bias']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
```

In [7]:
```
# model_checkpoint = 'distilbert-base-uncased'
# tokenizer = AutoTokenizer.from_pretrained(model_checkpoint, use_fast=False)
# model = AutoModelForSequenceClassification.from_pretrained(model_checkpoint, num_labels=2).to('cuda')
```

In [8]:
```
import re
```

In [9]:
```
def remove_URL(text):
    url = re.compile(r'https?://\S+|www\.\S+')
    return url.sub(r'',text)
```

In [10]:
```
def remove_html(text):
    html=re.compile(r'<.*?>')
    return html.sub(r'',text)

example = """<div>
<h1>Real or Fake</h1>
<p>Kaggle </p>
<a href="https://www.kaggle.com/c/nlp-getting-started">getting started</a>
</div>"""
print(remove_html(example))
```

```
Real or Fake
Kaggle
getting started
```

In [11]:
```
# Reference : https://gist.github.com/slowkow/7a7f61f495e3dbb7e3d767f97bd7304b
def remove_emoji(text):
    emoji_pattern = re.compile("["
                           u"\U0001F600-\U0001F64F"  # emoticons
                           u"\U0001F300-\U0001F5FF"  # symbols & pictographs
                           u"\U0001F680-\U0001F6FF"  # transport & map symbols
                           u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                           u"\U00002702-\U000027B0"
                           u"\U000024C2-\U0001F251"
                           "]+", flags=re.UNICODE)
    return emoji_pattern.sub(r'', text)

remove_emoji("Omg another Earthquake 😔😔")
```

Out[11]:
```
'Omg another Earthquake '
```

In [12]:
```
def remove(text_list):
    for i in range(len(text_list)):
        text_list[i] = remove_URL(text_list[i])
        text_list[i] = remove_html(text_list[i])
        text_list[i] = remove_emoji(text_list[i])


    print(text_list)
    return text_list
```

In [13]:
```
def preprocess_function(examples):
    return tokenizer(examples['text'], truncation=True)
    # return tokenizer(remove(examples['text']), truncation=True)
```

In [14]:
```
train_data = Dataset.from_pandas(train_data)
dev_data = Dataset.from_pandas(dev_data)
test_data = Dataset.from_pandas(test_data)

encoded_dataset_train = train_data.map(preprocess_function, batched=True)
encoded_dataset_dev = dev_data.map(preprocess_function, batched=True)
encoded_dataset_test = test_data.map(preprocess_function, batched=True)
```

```
  0%|          | 0/7 [00:00<?, ?ba/s]

  0%|          | 0/1 [00:00<?, ?ba/s]

  0%|          | 0/8 [00:00<?, ?ba/s]
```

In [15]:
```
columns_to_return = ['input_ids', 'label', 'attention_mask']
columns_to_return_test = ['input_ids', 'attention_mask']
encoded_dataset_train.set_format(columns=columns_to_return)
encoded_dataset_dev.set_format(columns=columns_to_return)
encoded_dataset_test.set_format(columns=columns_to_return_test)
```

In [16]:
```
# batch_size = 8
# batch_size = 16
batch_size = 32
metric_name = "f1"
model_name = model_checkpoint.split("/")[-1]
task = 'tweet'

args = TrainingArguments(
    f"./save_model/{model_name}-finetuned-{task}",
    evaluation_strategy = "epoch",
    save_strategy = "epoch",
    learning_rate=1e-5,
    per_device_train_batch_size=batch_size,
    per_device_eval_batch_size=batch_size,
    num_train_epochs=3,
    weight_decay=0.01,
    load_best_model_at_end=True,
    metric_for_best_model=metric_name,
    push_to_hub=False,
)
```

In [17]:
```
metric = load_metric('f1')
def compute_metrics(eval_pred):
    predictions, labels = eval_pred
    predictions = np.argmax(predictions, axis=1)
    return metric.compute(predictions=predictions, references=labels, average='macro')
```

```
Downloading builder script:   0%|          | 0.00/2.32k [00:00<?, ?B/s]
```

In [18]:
```
trainer = Trainer(
    model,
    args,
    train_dataset=encoded_dataset_train,
    eval_dataset=encoded_dataset_dev,
    tokenizer=tokenizer,
    compute_metrics=compute_metrics
)
```

In [19]:
```
trainer.train()
```

```
The following columns in the training set don't have a corresponding argument in `RobertaForSequenceClassification.forward` and have been ignored: text, __index_level_0__. If text, __index_level_0__ are not expected by `RobertaForSequenceClassification.forward`,  you can safely ignore this message.
/usr/local/lib/python3.7/dist-packages/transformers/optimization.py:309: FutureWarning: This implementation of AdamW is deprecated and will be removed in a future version. Use the PyTorch implementation torch.optim.AdamW instead, or set `no_deprecation_warning=True` to disable this warning
  FutureWarning,
***** Running training *****
  Num examples = 6851
  Num Epochs = 3
  Instantaneous batch size per device = 32
  Total train batch size (w. parallel, distributed & accumulation) = 32
  Gradient Accumulation steps = 1
  Total optimization steps = 645
```


[645/645 13:23, Epoch 3/3]

Epoch	Training Loss	Validation Loss	F1
1	No log	0.394217	0.827980
2	No log	0.395884	0.832316
3	0.394400	0.423998	0.832392

```
The following columns in the evaluation set don't have a corresponding argument in `RobertaForSequenceClassification.forward` and have been ignored: text, __index_level_0__. If text, __index_level_0__ are not expected by `RobertaForSequenceClassification.forward`,  you can safely ignore this message.
***** Running Evaluation *****
  Num examples = 762
  Batch size = 32
Saving model checkpoint to ./save_model/roberta-large-finetuned-tweet/checkpoint-215
Configuration saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-215/config.json
Model weights saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-215/pytorch_model.bin
tokenizer config file saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-215/tokenizer_config.json
Special tokens file saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-215/special_tokens_map.json
The following columns in the evaluation set don't have a corresponding argument in `RobertaForSequenceClassification.forward` and have been ignored: text, __index_level_0__. If text, __index_level_0__ are not expected by `RobertaForSequenceClassification.forward`,  you can safely ignore this message.
***** Running Evaluation *****
  Num examples = 762
  Batch size = 32
Saving model checkpoint to ./save_model/roberta-large-finetuned-tweet/checkpoint-430
Configuration saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-430/config.json
Model weights saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-430/pytorch_model.bin
tokenizer config file saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-430/tokenizer_config.json
Special tokens file saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-430/special_tokens_map.json
The following columns in the evaluation set don't have a corresponding argument in `RobertaForSequenceClassification.forward` and have been ignored: text, __index_level_0__. If text, __index_level_0__ are not expected by `RobertaForSequenceClassification.forward`,  you can safely ignore this message.
***** Running Evaluation *****
  Num examples = 762
  Batch size = 32
Saving model checkpoint to ./save_model/roberta-large-finetuned-tweet/checkpoint-645
Configuration saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-645/config.json
Model weights saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-645/pytorch_model.bin
tokenizer config file saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-645/tokenizer_config.json
Special tokens file saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-645/special_tokens_map.json


Training completed. Do not forget to share your model on huggingface.co/models =)


Loading best model from ./save_model/roberta-large-finetuned-tweet/checkpoint-645 (score: 0.8323920077198161).
```

Out[19]:
```
TrainOutput(global_step=645, training_loss=0.3719262825426205, metrics={'train_runtime': 805.5417, 'train_samples_per_second': 25.515, 'train_steps_per_second': 0.801, 'total_flos': 2211928169788344.0, 'train_loss': 0.3719262825426205, 'epoch': 3.0})
```

In [20]:
```
#get_test_predictions
predictions = trainer.predict(encoded_dataset_test)
preds = np.argmax(predictions.predictions, axis=-1)
#print the top 100 examples
for i in range(100):
    print(encoded_dataset_test['text'][i], preds[i], sep='\t')
```

```
The following columns in the test set don't have a corresponding argument in `RobertaForSequenceClassification.forward` and have been ignored: text. If t
ext are not expected by `RobertaForSequenceClassification.forward`,  you can safely ignore this message.
***** Running Prediction *****
  Num examples = 7613
  Batch size = 32
```

[238/238 01:22]

```
Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all   1
Forest fire near La Ronge Sask. Canada  1
All residents asked to 'shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected   1
13,000 people receive #wildfires evacuation orders in California        1
Just got sent this photo from Ruby #Alaska as smoke from #wildfires pours into a school         1
#RockyFire Update => California Hwy. 20 closed in both directions due to Lake County fire - #CAfire #wildfires  1
#flood #disaster Heavy rain causes flash flooding of streets in Manitou, Colorado Springs areas 1
I'm on top of the hill and I can see a fire in the woods...      1
There's an emergency evacuation happening now in the building across the street 1
I'm afraid that the tornado is coming to our area...     1
Three people died from the heat wave so far      1
Haha South Tampa is getting flooded hah- WAIT A SECOND I LIVE IN SOUTH TAMPA WHAT AM I GONNA DO WHAT AM I GONNA DO FVCK #flooding       1
#raining #flooding #Florida #TampaBay #Tampa 18 or 19 days. I've lost count      1
#Flood in Bago Myanmar #We arrived Bago 1
Damage to school bus on 80 in multi car crash #BREAKING          1
What's up man?  0
I love fruits   0
Summer is lovely        0
My car is so fast       0
What a goooooooaaaaaal!!!!!!    0
this is ridiculous....  0
London is cool ;)       0
Love skiing     0
What a wonderful day!   0
LOOOOOOL        0
No way...I can't eat that shit  0
Was in NYC last week!   0
Love my girlfriend      0
Cooool :)       0
Do you like pasta?      0
The end!        0
@bbcmtd Wholesale Markets ablaze http://t.co/lHYXEOHY6C (http://t.co/lHYXEOHY6C)         1
We always try to bring the heavy. #metal #RT http://t.co/YAo1eOxngw (http://t.co/YAo1eOxngw)     0
#AFRICANBAZE: Breaking news:Nigeria flag set ablaze in Aba. http://t.co/2nndBGwyEi (http://t.co/2nndBGwyEi)      1
Crying out for more! Set me ablaze      0
On plus side LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE http://t.co/qqsmshaJ3N (http://t.co/qqsmshaJ3N)    1
@PhDSquares #mufc they've built so much hype around new acquisitions but I doubt they will set the EPL ablaze this season.       0
INEC Office in Abia Set Ablaze - http://t.co/3ImaomknnA (http://t.co/3ImaomknnA)         1
Barbados #Bridgetown JAMAICA   ÛÒ Two cars set ablaze: SANTA CRUZ   ÛÓ Head of the St Elizabeth Police Superintende...  http://t.co/wDUEaj8Q4J (http://t.co/wDUEaj8Q4J)   1
Ablaze for you Lord :D  0
Check these out: http://t.co/rOI2NSmEJJ (http://t.co/rOI2NSmEJJ) http://t.co/3Tj8ZjiN21 (http://t.co/3Tj8ZjiN21) http://t.co/YDUiXEfIpE (http://t.co/YDUiXEfIpE) http://t.co/LxTjc87KLS (http://t.co/LxTjc87KLS) #nsfw   0
on the outside you're ablaze and alive
but you're dead inside  1
Had an awesome time visiting the CFC head office the ancop site and ablaze. Thanks to Tita Vida for taking care of us ??         0
SOOOO PUMPED FOR ABLAZE ???? @southridgelife    0
I wanted to set Chicago ablaze with my preaching... But not my hotel! http://t.co/o9qknbfOFX (http://t.co/o9qknbfOFX)    0
I gained 3 followers in the last week. You? Know your stats and grow with http://t.co/TIyUliF5c6 (http://t.co/TIyUliF5c6)       0
How the West was burned: Thousands of wildfires ablaze in California alone http://t.co/vl5TBR3wbr (http://t.co/vl5TBR3wbr)      1
Building the perfect tracklist to life leave the streets ablaze 0
Check these out: http://t.co/rOI2NSmEJJ (http://t.co/rOI2NSmEJJ) http://t.co/3Tj8ZjiN21 (http://t.co/3Tj8ZjiN21) http://t.co/YDUiXEfIpE (http://t.co/YDUiXEfIpE) http://t.co/LxTjc87KLS (http://t.co/LxTjc87KLS) #nsfw   0
First night with retainers in. It's quite weird. Better get used to it; I have to wear them every single night for the next year at least.      0
Deputies: Man shot before Brighton home set ablaze http://t.co/gWNRhMSO8k (http://t.co/gWNRhMSO8k)      1
Man wife get six years jail for setting ablaze niece
http://t.co/eV1ahOUCZA (http://t.co/eV1ahOUCZA) 1
SANTA CRUZ   ÛÓ Head of the St Elizabeth Police Superintendent Lanford Salmon has r ... - http://t.co/vplR5Hka2u (http://t.co/vplR5Hka2u) http://t.co/SxHW2TNNLf (http://t.co/SxHW2TNNLf)         0
Police: Arsonist Deliberately Set Black Church In North CarolinaåÊAblaze http://t.co/pcXarbH9An (http://t.co/pcXarbH9An)         1
Noches El-Bestia '@Alexis_Sanchez: happy to see my teammates and training hard ?? goodnight gunners.?????? http://t.co/uc4j4jHvGR' (http://t.co/uc4j4jHvGR')     0
#Kurds trampling on Turkmen flag later set it ablaze while others vandalized offices of Turkmen Front in #Diyala http://t.co/4IzFdYC3cg (http://t.co/4IzFdYC3cg) 1
TRUCK ABLAZE : R21. VOORTREKKER AVE. OUTSIDE OR TAMBO INTL. CARGO SECTION. http://t.co/8kscqKfKkF (http://t.co/8kscqKfKkF)      1
Set our hearts ablaze and every city was a gift And every skyline was like a kiss upon the lips @  Û_ https://t.co/cYoMPZ1AOZ (https://t.co/cYoMPZ1AOZ) 0
They sky was ablaze tonight in Los Angeles. I'm expecting IG and FB to be filled with sunset shots if I know my peeps!! 0
How the West was burned: Thousands of wildfires ablaze in #California alone http://t.co/iCSjGZ9tE1 (http://t.co/iCSjGZ9tE1) #climate #energy http://t.co/9FxmNOlOBd (http://t.co/9FxmNOlOBd)     1
Revel in yours wmv videos by means of mac farewell ablaze wmv en route to dvd: GtxRWm   0
Progressive greetings!

In about a month students would have set their pens ablaze in The Torch Publications'... http://t.co/9FxPiXQuJt (http://t.co/9FxPiXQuJt)         0
Rene Ablaze &amp;amp; Jacinta - Secret 2k13 (Fallen Skies Edit) - Mar 30 2013  https://t.co/7MLMsUzV1Z (https://t.co/7MLMsUzV1Z)     0
@Navista7 Steve these fires out here are something else! California is a tinderbox - and this clown was setting my 'hood ablaze @News24680      1
#NowPlaying: Rene Ablaze &amp;amp; Ian Buff - Magnitude http://t.co/Av2JSjfFtc (http://t.co/Av2JSjfFtc)  #EDM       0
@nxwestmidlands huge fire at Wholesale markets ablaze http://t.co/rwzbFVNXER (http://t.co/rwzbFVNXER)   1
@ablaze what time does your talk go until? I don't know if I can make it due to work.   0
'I can't have kids cuz I got in a bicycle accident &amp;amp; split my testicles. it's impossible for me to have kids' MICHAEL YOU ARE THE FATHER    0
Accident on I-24 W #NashvilleTraffic. Traffic moving 8m slower than usual. https://t.co/OGHk693EgJ (https://t.co/OGHk693EgJ)     1
Accident center lane blocked in #SantaClara on US-101 NB before Great America Pkwy #BayArea #Traffic http://t.co/pmlOhZuRWR (http://t.co/pmlOhZuRWR)     1
http://t.co/GKYe6gjTk5 (http://t.co/GKYe6gjTk5) Had a #personalinjury accident this summer? Read our advice &amp;amp; see how a #solicitor can help #OtleyHour       0
#stlouis #caraccidentlawyer Speeding Among Top Causes of Teen Accidents https://t.co/k4zoMOF319 (https://t.co/k4zoMOF319) https://t.co/S2kXVMOcBA (https://t.co/S2kXVMOcBA) Car Accident tee  Û_       1
Reported motor vehicle accident in Curry on Herman Rd near Stephenson involving an overturned vehicle. Please use... http://t.co/YbJezKuRW1 (http://t.co/
```

```
YbJezKuRW1)     1
BigRigRadio Live Accident Awareness     1
I-77 Mile Marker 31 South Mooresville  Iredell Vehicle Accident Ramp Closed at 8/6 1:18 PM      1
RT @SleepJunkies: Sleeping pills double your risk of a car accident http://t.co/7s9Nm1fiCT (http://t.co/7s9Nm1fiCT)     0
'By accident' they knew what was gon happen https://t.co/Ysxun5vCeh (https://t.co/Ysxun5vCeh)    0
Traffic accident N CABRILLO HWY/MAGELLAN AV MIR (08/06/15 11:03:58)     1
I-77 Mile Marker 31 to 40 South Mooresville  Iredell Vehicle Accident Congestion at 8/6 1:18 PM 1
the pastor was not in the scene of the accident.....who was the owner of the range rover ?     1
mom: 'we didn't get home as fast as we wished'
me: 'why is that?'
mom: 'there was an accident and some truck spilt mayonnaise all over ??????     0
I was in a horrible car accident this past Sunday. I'm finally able to get around. Thank you GOD??      1
Can wait to see how pissed Donnie is when I tell him I was in ANOTHER accident??        0
#TruckCrash Overturns On #FortWorth Interstate http://t.co/Rs22LJ4qFp (http://t.co/Rs22LJ4qFp) Click here if you've been in a crash&gt;http://t.co/LdOunI
Yw4k    1
Accident in #Ashville on US 23 SB before SR 752 #traffic http://t.co/hylMoOWgFI (http://t.co/hylMoOWgFI)        1
Carolina accident: Motorcyclist Dies in I-540 Crash With Car That Crossed Median: A motorcycle rider traveling... http://t.co/p18lzRlmy6 (http://t.co/p18
lzRlmy6)        1
FYI CAD:FYI: ;ACCIDENT PROPERTY DAMAGE;NHS;999 PINER RD/HORNDALE DR     1
RT nAAYf: First accident in years. Turning onto Chandanee Magu from near MMA. Taxi rammed into me while I was halfway turned. Everyone conf  Û_ 1
Accident left lane blocked in #Manchester on Rt 293 NB before Eddy Rd stop and go traffic back to NH-3A delay of 4 mins #traffic        1
;ACCIDENT PROPERTY DAMAGE; PINER RD/HORNDALE DR 1
???? it was an accident http://t.co/Oia5fxi4gM (http://t.co/Oia5fxi4gM) 0
FYI CAD:FYI: ;ACCIDENT PROPERTY DAMAGE;WPD;1600 S 17TH ST       1
8/6/2015@2:09 PM: TRAFFIC ACCIDENT NO INJURY at 2781 WILLIS FOREMAN RD http://t.co/VCkIT6EDEv (http://t.co/VCkIT6EDEv)  1
Aashiqui Actress Anu Aggarwal On Her Near-Fatal Accident http://t.co/6Otfp31LqW (http://t.co/6Otfp31LqW)        0
Suffield Alberta Accident https://t.co/bPTmlF4P1O (https://t.co/bPTmlF4P1O)     1
9 Mile backup on I-77 South...accident blocking the Right 2 Lanes at Exit 31 Langtree Rd...consider NC 115 or NC 150 to NC 16 as alternate      1
Has an accident changed your life? We will help you determine options that can financially support life care plans and on-going treatment.      0
#BREAKING: there was a deadly motorcycle car accident that happened to #Hagerstown today. I'll have more details at 5 @Your4State. #WHAG       1
@flowri were you marinading it or was it an accident?   0
only had a car for not even a week and got in a fucking car accident .. Mfs can't fucking drive .       0
```

In [20]:

Optional: custom class weight

In [21]:
```
train_labels = encoded_dataset_train['label']
print(np.bincount(train_labels))
class_weights = compute_class_weight(class_weight='balanced', classes=np.unique(train_labels), y=list(train_labels))
print(class_weights)
```

```
[3907 2944]
[0.87675966 1.16355299]
```

In [22]:
```
class CustomTrainer(Trainer):
    def compute_loss(self, model, inputs, return_outputs=False):
        #print(inputs)
        labels = inputs.get("labels")
        # forward pass
        outputs = model(**inputs)
        logits = outputs.get("logits")
        # compute custom loss (suppose one has 2 labels with different weights)
        loss_fct = nn.CrossEntropyLoss(weight=torch.Tensor(class_weights).to('cuda'))
        loss = loss_fct(logits.view(-1, self.model.config.num_labels), labels.view(-1))
        return (loss, outputs) if return_outputs else loss
```

In [23]:
```
trainer = CustomTrainer(
    model,
    args,
    train_dataset=encoded_dataset_train,
    eval_dataset=encoded_dataset_dev,
    tokenizer=tokenizer,
    compute_metrics=compute_metrics
)
```

In [24]:
```
trainer.train()
```

```
The following columns in the training set don't have a corresponding argument in `RobertaForSequenceClassification.forward` and have been ignored: text,
__index_level_0__. If text, __index_level_0__ are not expected by `RobertaForSequenceClassification.forward`,  you can safely ignore this message.
/usr/local/lib/python3.7/dist-packages/transformers/optimization.py:309: FutureWarning: This implementation of AdamW is deprecated and will be removed in
a future version. Use the PyTorch implementation torch.optim.AdamW instead, or set `no_deprecation_warning=True` to disable this warning
  FutureWarning,
***** Running training *****
  Num examples = 6851
  Num Epochs = 3
  Instantaneous batch size per device = 32
  Total train batch size (w. parallel, distributed & accumulation) = 32
  Gradient Accumulation steps = 1
  Total optimization steps = 645
```

[645/645 13:25, Epoch 3/3]

Epoch	Training Loss	Validation Loss	F1
1	No log	0.475748	0.832343
2	No log	0.543113	0.823952
3	0.210000	0.633628	0.822806

```
The following columns in the evaluation set don't have a corresponding argument in `RobertaForSequenceClassification.forward` and have been ignored: tex
t, __index_level_0__. If text, __index_level_0__ are not expected by `RobertaForSequenceClassification.forward`,  you can safely ignore this message.
***** Running Evaluation *****
  Num examples = 762
  Batch size = 32
Saving model checkpoint to ./save_model/roberta-large-finetuned-tweet/checkpoint-215
Configuration saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-215/config.json
Model weights saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-215/pytorch_model.bin
tokenizer config file saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-215/tokenizer_config.json
Special tokens file saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-215/special_tokens_map.json
The following columns in the evaluation set don't have a corresponding argument in `RobertaForSequenceClassification.forward` and have been ignored: tex
t, __index_level_0__. If text, __index_level_0__ are not expected by `RobertaForSequenceClassification.forward`,  you can safely ignore this message.
***** Running Evaluation *****
  Num examples = 762
  Batch size = 32
Saving model checkpoint to ./save_model/roberta-large-finetuned-tweet/checkpoint-430
Configuration saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-430/config.json
Model weights saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-430/pytorch_model.bin
tokenizer config file saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-430/tokenizer_config.json
Special tokens file saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-430/special_tokens_map.json
The following columns in the evaluation set don't have a corresponding argument in `RobertaForSequenceClassification.forward` and have been ignored: tex
t, __index_level_0__. If text, __index_level_0__ are not expected by `RobertaForSequenceClassification.forward`,  you can safely ignore this message.
***** Running Evaluation *****
  Num examples = 762
  Batch size = 32
Saving model checkpoint to ./save_model/roberta-large-finetuned-tweet/checkpoint-645
Configuration saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-645/config.json
Model weights saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-645/pytorch_model.bin
tokenizer config file saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-645/tokenizer_config.json
Special tokens file saved in ./save_model/roberta-large-finetuned-tweet/checkpoint-645/special_tokens_map.json


Training completed. Do not forget to share your model on huggingface.co/models =)


Loading best model from ./save_model/roberta-large-finetuned-tweet/checkpoint-215 (score: 0.8323432343234324).
```

Out[24]:
```
TrainOutput(global_step=645, training_loss=0.20530634443889292, metrics={'train_runtime': 807.1734, 'train_samples_per_second': 25.463, 'train_steps_per_
second': 0.799, 'total_flos': 2211928169788344.0, 'train_loss': 0.20530634443889292, 'epoch': 3.0})
```

In [24]:
```
```