

```
In [1]: !pip install transformers
!pip install datasets
!pip3 install emoji==0.6.0
```

```
Requirement already satisfied: transformers in /usr/local/lib/python3.7/dist-packages (4.19.1)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.7/dist-packages (from transformers) (21.3)
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)
Requirement already satisfied: pyyaml!=5.1 in /usr/local/lib/python3.7/dist-packages (from transformers) (6.0)
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)
Requirement already satisfied: huggingface-hub<1.0,>=0.1.0 in /usr/local/lib/python3.7/dist-packages (from transformers) (0.6.0)
Requirement already satisfied: tokenizers!=0.11.3,<0.13,>=0.11.1 in /usr/local/lib/python3.7/dist-packages (from transformers) (0.12.1)
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from transformers) (4.11.3)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-packages (from transformers) (1.21.6)
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: urllib3!=1.25.0,!1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-packages (from requests->transformers) (1.25.11)
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: datasets in /usr/local/lib/python3.7/dist-packages (2.2.1)
Requirement already satisfied: fsspec[http]>=2021.05.0 in /usr/local/lib/python3.7/dist-packages (from datasets) (2022.3.0)
Requirement already satisfied: pandas in /usr/local/lib/python3.7/dist-packages (from datasets) (1.3.5)
Requirement already satisfied: pyarrow>=6.0.0 in /usr/local/lib/python3.7/dist-packages (from datasets) (6.0.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.7/dist-packages (from datasets) (21.3)
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from datasets) (4.11.3)
Requirement already satisfied: multiprocessing in /usr/local/lib/python3.7/dist-packages (from datasets) (0.70.12.2)
Requirement already satisfied: requests>=2.19.0 in /usr/local/lib/python3.7/dist-packages (from datasets) (2.23.0)
Requirement already satisfied: xxhash in /usr/local/lib/python3.7/dist-packages (from datasets) (3.0.0)
Requirement already satisfied: dill in /usr/local/lib/python3.7/dist-packages (from datasets) (0.3.4)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-packages (from datasets) (1.21.6)
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: tqdm>=4.62.1 in /usr/local/lib/python3.7/dist-packages (from datasets) (4.64.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.7/dist-packages (from datasets) (3.8.1)
Requirement already satisfied: responses<0.19 in /usr/local/lib/python3.7/dist-packages (from datasets) (0.18.0)
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: pyyaml in /usr/local/lib/python3.7/dist-packages (from huggingface-hub<1.0.0,>=0.1.0->datasets) (6.0)
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: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests>=2.19.0->datasets) (2.10)
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: 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) (1.25.11)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.7/dist-packages (from aiohttp->datasets) (6.0.2)
Requirement already satisfied: charset-normalizer<3.0,>=2.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp->datasets) (2.0.12)
Requirement already satisfied: async-timeout<5.0,>=4.0.0a3 in /usr/local/lib/python3.7/dist-packages (from aiohttp->datasets) (4.0.2)
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp->datasets) (1.7.2)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.7/dist-packages (from aiohttp->datasets) (1.2.0)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp->datasets) (21.4.0)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metadata->datasets) (3.8.0)
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: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from pandas->datasets) (2022.1)
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)
Requirement already satisfied: emoji==0.6.0 in /usr/local/lib/python3.7/dist-packages (0.6.0)
```

```
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, AutoModel
from transformers import AutoModelForSequenceClassification, TrainingArguments, Trainer, DataCollatorWithPadding
```

```
In [3]: 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 [4]: train_path = '/content/drive/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/drive/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)

train_data.head()
```

```
Out[4]:
```

	text	label
0	Our Deeds are the Reason of this #earthquake M...	1
1	Forest fire near La Ronge Sask. Canada	1
2	All residents asked to 'shelter in place' are ...	1
3	13,000 people receive #wildfires evacuation or...	1
4	Just got sent this photo from Ruby #Alaska as ...	1

```
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']))
```

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

```
In [6]: # BERTweet provides this pre-processing step by enabling the normalization argument of its tokenizer.
# Therefore, there is no separate pre-processing code for this model
bertweet = "vinai/bertweet-base"
tokenizer = AutoTokenizer.from_pretrained(bertweet,
                                         use_fast=False,
                                         normalization=True)

model = AutoModelForSequenceClassification.from_pretrained(
    bertweet,
    num_labels = 2,
    output_attentions = False,
    output_hidden_states = False,
)
```

Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.

Some weights of the model checkpoint at vinai/bertweet-base were not used when initializing RobertaForSequenceClassification: ['roberta.pooler.dense.bias', 'lm_head.bias', 'lm_head.decoder.bias', 'lm_head.layer_norm.weight', 'lm_head.decoder.weight', 'lm_head.layer_norm.bias', 'lm_head.dense.weight', 'roberta.pooler.dense.weight', 'lm_head.dense.bias']

- This IS expected if you are initializing RobertaForSequenceClassification from the checkpoint of a model trained on another task or with another architecture (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 (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).

Some weights of RobertaForSequenceClassification were not initialized from the model checkpoint at vinai/bertweet-base and are newly initialized: ['classifier.out_proj.weight', 'classifier.dense.weight', 'classifier.out_proj.bias', 'classifier.dense.bias']

You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

```
In [7]: def preprocess_function(examples):
        return tokenizer(examples['text'], truncation=True)
```

```
In [8]: 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 [9]: 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 [10]: batch_size = 80
metric_name = "f1"
model_name = bertweet.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=5,
    weight_decay=0.01,
    load_best_model_at_end=True,
    metric_for_best_model=metric_name,
    push_to_hub=False,
)
```

```
In [11]: 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')
```

```
In [12]: trainer = Trainer(
    model,
    args,
    train_dataset=encoded_dataset_train,
    eval_dataset=encoded_dataset_dev,
    tokenizer=tokenizer,
    compute_metrics=compute_metrics
)
```

```
In [13]: 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 = 5
Instantaneous batch size per device = 80
Total train batch size (w. parallel, distributed & accumulation) = 80
Gradient Accumulation steps = 1
Total optimization steps = 430

[430/430 04:40, Epoch 5/5]

Epoch	Training Loss	Validation Loss	F1
1	No log	0.373162	0.856727
2	No log	0.358687	0.855437
3	No log	0.355969	0.859151
4	No log	0.356055	0.858015
5	No log	0.358279	0.858607

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 = 80
Saving model checkpoint to ./save_model/bertweet-base-finetuned-tweet/checkpoint-86
Configuration saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-86/config.json
Model weights saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-86/pytorch_model.bin
tokenizer config file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-86/tokenizer_config.json
Special tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-86/special_tokens_map.json
added tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-86/added_tokens.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 = 80
Saving model checkpoint to ./save_model/bertweet-base-finetuned-tweet/checkpoint-172
Configuration saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-172/config.json
Model weights saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-172/pytorch_model.bin
tokenizer config file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-172/tokenizer_config.json
Special tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-172/special_tokens_map.json
added tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-172/added_tokens.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 = 80
Saving model checkpoint to ./save_model/bertweet-base-finetuned-tweet/checkpoint-258
Configuration saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-258/config.json
Model weights saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-258/pytorch_model.bin
tokenizer config file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-258/tokenizer_config.json
Special tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-258/special_tokens_map.json
added tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-258/added_tokens.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 = 80
Saving model checkpoint to ./save_model/bertweet-base-finetuned-tweet/checkpoint-344
Configuration saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-344/config.json
Model weights saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-344/pytorch_model.bin
tokenizer config file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-344/tokenizer_config.json
Special tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-344/special_tokens_map.json
added tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-344/added_tokens.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 = 80
Saving model checkpoint to ./save_model/bertweet-base-finetuned-tweet/checkpoint-430
Configuration saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-430/config.json
Model weights saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-430/pytorch_model.bin
tokenizer config file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-430/tokenizer_config.json
Special tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-430/special_tokens_map.json
added tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-430/added_tokens.json

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

Loading best model from ./save_model/bertweet-base-finetuned-tweet/checkpoint-258 (score: 0.8591509865676061).

```
Out[13]: TrainOutput(global_step=430, training_loss=0.3870725764784702, metrics={'train_runtime': 281.2095, 'train_samples_per_second': 121.813, 'train_steps_per_second': 1.529, 'total_flos': 716950389750660.0, 'train_loss': 0.3870725764784702, 'epoch': 5.0})
```

```
In [14]: #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 text are not expected by `RobertaForSequenceClassification.forward`, you can safely ignore this message.

**** Running Prediction ****

Num examples = 7613

Batch size = 80

[96/96 00:17]

Optional: custom class weight

```
In [15]: 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 [16]: 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 [17]: trainer = CustomTrainer(
    model,
    args,
    train_dataset=encoded_dataset_train,
    eval_dataset=encoded_dataset_dev,
    tokenizer=tokenizer,
    compute_metrics=compute_metrics
)
```

```
In [18]: 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 = 5
Instantaneous batch size per device = 80
Total train batch size (w. parallel, distributed & accumulation) = 80
Gradient Accumulation steps = 1
Total optimization steps = 430

[430/430 04:41, Epoch 5/5]

Epoch	Training Loss	Validation Loss	F1
1	No log	0.372555	0.853745
2	No log	0.374509	0.855697
3	No log	0.386352	0.860713
4	No log	0.391802	0.859026
5	No log	0.399911	0.853502

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 = 80
Saving model checkpoint to ./save_model/bertweet-base-finetuned-tweet/checkpoint-86
Configuration saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-86/config.json
Model weights saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-86/pytorch_model.bin
tokenizer config file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-86/tokenizer_config.json
Special tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-86/special_tokens_map.json
added tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-86/added_tokens.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 = 80
Saving model checkpoint to ./save_model/bertweet-base-finetuned-tweet/checkpoint-172
Configuration saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-172/config.json
Model weights saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-172/pytorch_model.bin
tokenizer config file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-172/tokenizer_config.json
Special tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-172/special_tokens_map.json
added tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-172/added_tokens.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 = 80
Saving model checkpoint to ./save_model/bertweet-base-finetuned-tweet/checkpoint-258
Configuration saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-258/config.json
Model weights saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-258/pytorch_model.bin
tokenizer config file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-258/tokenizer_config.json
Special tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-258/special_tokens_map.json
added tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-258/added_tokens.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 = 80
Saving model checkpoint to ./save_model/bertweet-base-finetuned-tweet/checkpoint-344
Configuration saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-344/config.json
Model weights saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-344/pytorch_model.bin
tokenizer config file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-344/tokenizer_config.json
Special tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-344/special_tokens_map.json
added tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-344/added_tokens.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 = 80
Saving model checkpoint to ./save_model/bertweet-base-finetuned-tweet/checkpoint-430
Configuration saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-430/config.json
Model weights saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-430/pytorch_model.bin
tokenizer config file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-430/tokenizer_config.json
Special tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-430/special_tokens_map.json
added tokens file saved in ./save_model/bertweet-base-finetuned-tweet/checkpoint-430/added_tokens.json

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

Loading best model from ./save_model/bertweet-base-finetuned-tweet/checkpoint-258 (score: 0.8607127455486678).

```
Out[18]: TrainOutput(global_step=430, training_loss=0.2853202642396439, metrics={'train_runtime': 282.1969, 'train_samples_per_second': 121.387, 'train_steps_per_second': 1.524, 'total_flos': 716950389750660.0, 'train_loss': 0.2853202642396439, 'epoch': 5.0})
```

```
In [19]: #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 text are not expected by `RobertaForSequenceClassification.forward`, you can safely ignore this message.

**** Running Prediction ****

Num examples = 7613

Batch size = 80

[96/96 00:17]