final-bert

August 15, 2024

```
[ ]: !pip install transformers
     !pip install torch
```

Requirement already satisfied: transformers in /usr/local/lib/python3.10/dist-
packages (4.41.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from transformers) (3.14.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.23.0 in
/usr/local/lib/python3.10/dist-packages (from transformers) (0.23.2)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-
packages (from transformers) (1.25.2)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from transformers) (24.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-
packages (from transformers) (6.0.1)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.10/dist-packages (from transformers) (2024.5.15)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-
packages (from transformers) (2.31.0)
Requirement already satisfied: tokenizers<0.20,>=0.19 in
/usr/local/lib/python3.10/dist-packages (from transformers) (0.19.1)
Requirement already satisfied: safetensors>=0.4.1 in
/usr/local/lib/python3.10/dist-packages (from transformers) (0.4.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-
packages (from transformers) (4.66.4)
Requirement already satisfied: fsspec>=2023.5.0 in
/usr/local/lib/python3.10/dist-packages (from huggingface-
hub<1.0,>=0.23.0->transformers) (2023.6.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface-
hub<1.0,>=0.23.0->transformers) (4.12.1)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests->transformers) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in

/usr/local/lib/python3.10/dist-packages (from requests->transformers) (2024.6.2)
Requirement already satisfied: torch in /usr/local/lib/python3.10/dist-packages
(2.3.0+cu121)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from torch) (3.14.0)
Requirement already satisfied: typing-extensions>=4.8.0 in
/usr/local/lib/python3.10/dist-packages (from torch) (4.12.1)
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages
(from torch) (1.12.1)
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-
packages (from torch) (3.3)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages
(from torch) (3.1.4)
Requirement already satisfied: fsspec in /usr/local/lib/python3.10/dist-packages
(from torch) (2023.6.0)
Collecting nvidia-cuda-nvrtc-cu12==12.1.105 (from torch)
  Using cached nvidia_cuda_nvrtc_cu12-12.1.105-py3-none-manylinux1_x86_64.whl
(23.7 MB)
Collecting nvidia-cuda-runtime-cu12==12.1.105 (from torch)
  Using cached nvidia_cuda_runtime_cu12-12.1.105-py3-none-manylinux1_x86_64.whl
(823 kB)
Collecting nvidia-cuda-cupti-cu12==12.1.105 (from torch)
  Using cached nvidia_cuda_cupti_cu12-12.1.105-py3-none-manylinux1_x86_64.whl
(14.1 MB)
Collecting nvidia-cudnn-cu12==8.9.2.26 (from torch)
  Using cached nvidia_cudnn_cu12-8.9.2.26-py3-none-manylinux1_x86_64.whl (731.7
MB)
Collecting nvidia-cublas-cu12==12.1.3.1 (from torch)
  Using cached nvidia_cublas_cu12-12.1.3.1-py3-none-manylinux1_x86_64.whl (410.6
MB)
Collecting nvidia-cufft-cu12==11.0.2.54 (from torch)
  Using cached nvidia_cufft_cu12-11.0.2.54-py3-none-manylinux1_x86_64.whl (121.6
MB)
Collecting nvidia-curand-cu12==10.3.2.106 (from torch)
  Using cached nvidia_curand_cu12-10.3.2.106-py3-none-manylinux1_x86_64.whl
(56.5 MB)
Collecting nvidia-cusolver-cu12==11.4.5.107 (from torch)
  Using cached nvidia_cusolver_cu12-11.4.5.107-py3-none-manylinux1_x86_64.whl
(124.2 MB)
Collecting nvidia-cusparse-cu12==12.1.0.106 (from torch)
  Using cached nvidia_cusparse_cu12-12.1.0.106-py3-none-manylinux1_x86_64.whl
(196.0 MB)
Collecting nvidia-nccl-cu12==2.20.5 (from torch)
  Using cached nvidia_nccl_cu12-2.20.5-py3-none-manylinux2014_x86_64.whl (176.2
MB)
Collecting nvidia-nvtx-cu12==12.1.105 (from torch)
  Using cached nvidia_nvtx_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (99 kB)
Requirement already satisfied: triton==2.3.0 in /usr/local/lib/python3.10/dist-

packages (from torch) (2.3.0)
Collecting nvidia-nvjitlink-cu12 (from nvidia-cusolver-cu12==11.4.5.107->torch)
  Downloading nvidia_nvjitlink_cu12-12.5.40-py3-none-manylinux2014_x86_64.whl
(21.3 MB)
                                    21.3/21.3 MB
52.9 MB/s eta 0:00:00
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from jinja2->torch) (2.1.5)
Requirement already satisfied: mpmath<1.4.0,>=1.1.0 in
/usr/local/lib/python3.10/dist-packages (from sympy->torch) (1.3.0)
Installing collected packages: nvidia-nvtx-cu12, nvidia-nvjitlink-cu12, nvidia-
nccl-cu12, nvidia-curand-cu12, nvidia-cufft-cu12, nvidia-cuda-runtime-cu12,
nvidia-cuda-nvrtc-cu12, nvidia-cuda-cupti-cu12, nvidia-cublas-cu12, nvidia-
cusparse-cu12, nvidia-cudnn-cu12, nvidia-cusolver-cu12
Successfully installed nvidia-cublas-cu12-12.1.3.1 nvidia-cuda-cupti-
cu12-12.1.105 nvidia-cuda-nvrtc-cu12-12.1.105 nvidia-cuda-runtime-cu12-12.1.105
nvidia-cudnn-cu12-8.9.2.26 nvidia-cufft-cu12-11.0.2.54 nvidia-curand-
cu12-10.3.2.106 nvidia-cusolver-cu12-11.4.5.107 nvidia-cusparse-cu12-12.1.0.106
nvidia-nccl-cu12-2.20.5 nvidia-nvjitlink-cu12-12.5.40 nvidia-nvtx-cu12-12.1.105

```
[ ]: !pip install datasets
     !pip install transformers[torch] accelerate -U
```

Collecting datasets
  Downloading datasets-2.19.2-py3-none-any.whl (542 kB)
                                    542.1/542.1
kB 8.3 MB/s eta 0:00:00
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from datasets) (3.14.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-
packages (from datasets) (1.25.2)
Requirement already satisfied: pyarrow>=12.0.0 in
/usr/local/lib/python3.10/dist-packages (from datasets) (14.0.2)
Requirement already satisfied: pyarrow-hotfix in /usr/local/lib/python3.10/dist-
packages (from datasets) (0.6)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Downloading dill-0.3.8-py3-none-any.whl (116 kB)
                                    116.3/116.3
kB 7.5 MB/s eta 0:00:00
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-
packages (from datasets) (2.0.3)
Collecting requests>=2.32.1 (from datasets)
  Downloading requests-2.32.3-py3-none-any.whl (64 kB)
                                    64.9/64.9 kB
3.5 MB/s eta 0:00:00
Requirement already satisfied: tqdm>=4.62.1 in
/usr/local/lib/python3.10/dist-packages (from datasets) (4.66.4)

3

```
Collecting xxhash (from datasets)
  Downloading
xxhash-3.4.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
                              194.1/194.1

kB 12.1 MB/s eta 0:00:00
Collecting multiprocess (from datasets)
  Downloading multiprocess-0.70.16-py310-none-any.whl (134 kB)
                              134.8/134.8

kB 10.0 MB/s eta 0:00:00
Requirement already satisfied: fsspec[http]<=2024.3.1,>=2023.1.0 in
/usr/local/lib/python3.10/dist-packages (from datasets) (2023.6.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.10/dist-
packages (from datasets) (3.9.5)
Requirement already satisfied: huggingface-hub>=0.21.2 in
/usr/local/lib/python3.10/dist-packages (from datasets) (0.23.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-
packages (from datasets) (24.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-
packages (from datasets) (6.0.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-
packages (from aiohttp->datasets) (23.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (6.0.5)
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.10/dist-
packages (from aiohttp->datasets) (1.9.4)
Requirement already satisfied: async-timeout<5.0,>=4.0 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (4.0.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.21.2->datasets)
(4.12.1)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.32.1->datasets)
(3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests>=2.32.1->datasets) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.32.1->datasets)
(2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.32.1->datasets)
(2024.6.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
```

/usr/local/lib/python3.10/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas->datasets) (2023.4)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-
packages (from pandas->datasets) (2024.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.8.2->pandas->datasets) (1.16.0)
Installing collected packages: xxhash, requests, dill, multiprocess, datasets
  Attempting uninstall: requests
    Found existing installation: requests 2.31.0
    Uninstalling requests-2.31.0:
      Successfully uninstalled requests-2.31.0
ERROR: pip's dependency resolver does not currently take into account all

the packages that are installed. This behaviour is the source of the following

dependency conflicts.

google-colab 1.0.0 requires requests==2.31.0, but you have requests 2.32.3 which

is incompatible.

Successfully installed datasets-2.19.2 dill-0.3.8 multiprocess-0.70.16
requests-2.32.3 xxhash-3.4.1
Requirement already satisfied: transformers[torch] in
/usr/local/lib/python3.10/dist-packages (4.41.2)
Collecting accelerate
  Downloading accelerate-0.31.0-py3-none-any.whl (309 kB)
                              309.4/309.4

kB 6.0 MB/s eta 0:00:00
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from transformers[torch]) (3.14.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.23.0 in
/usr/local/lib/python3.10/dist-packages (from transformers[torch]) (0.23.2)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-
packages (from transformers[torch]) (1.25.2)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from transformers[torch]) (24.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-
packages (from transformers[torch]) (6.0.1)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.10/dist-packages (from transformers[torch]) (2024.5.15)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-
packages (from transformers[torch]) (2.32.3)
Requirement already satisfied: tokenizers<0.20,>=0.19 in
/usr/local/lib/python3.10/dist-packages (from transformers[torch]) (0.19.1)
Requirement already satisfied: safetensors>=0.4.1 in
/usr/local/lib/python3.10/dist-packages (from transformers[torch]) (0.4.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-

packages (from transformers[torch]) (4.66.4)
Requirement already satisfied: torch in /usr/local/lib/python3.10/dist-packages
(from transformers[torch]) (2.3.0+cu121)
Requirement already satisfied: psutil in /usr/local/lib/python3.10/dist-packages
(from accelerate) (5.9.5)
Requirement already satisfied: fsspec>=2023.5.0 in
/usr/local/lib/python3.10/dist-packages (from huggingface-
hub<1.0,>=0.23.0->transformers[torch]) (2023.6.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface-
hub<1.0,>=0.23.0->transformers[torch]) (4.12.1)
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages
(from torch->transformers[torch]) (1.12.1)
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-
packages (from torch->transformers[torch]) (3.3)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages
(from torch->transformers[torch]) (3.1.4)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.1.105 in
/usr/local/lib/python3.10/dist-packages (from torch->transformers[torch])
(12.1.105)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.1.105 in
/usr/local/lib/python3.10/dist-packages (from torch->transformers[torch])
(12.1.105)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.1.105 in
/usr/local/lib/python3.10/dist-packages (from torch->transformers[torch])
(12.1.105)
Requirement already satisfied: nvidia-cudnn-cu12==8.9.2.26 in
/usr/local/lib/python3.10/dist-packages (from torch->transformers[torch])
(8.9.2.26)
Requirement already satisfied: nvidia-cublas-cu12==12.1.3.1 in
/usr/local/lib/python3.10/dist-packages (from torch->transformers[torch])
(12.1.3.1)
Requirement already satisfied: nvidia-cufft-cu12==11.0.2.54 in
/usr/local/lib/python3.10/dist-packages (from torch->transformers[torch])
(11.0.2.54)
Requirement already satisfied: nvidia-curand-cu12==10.3.2.106 in
/usr/local/lib/python3.10/dist-packages (from torch->transformers[torch])
(10.3.2.106)
Requirement already satisfied: nvidia-cusolver-cu12==11.4.5.107 in
/usr/local/lib/python3.10/dist-packages (from torch->transformers[torch])
(11.4.5.107)
Requirement already satisfied: nvidia-cusparse-cu12==12.1.0.106 in
/usr/local/lib/python3.10/dist-packages (from torch->transformers[torch])
(12.1.0.106)
Requirement already satisfied: nvidia-nccl-cu12==2.20.5 in
/usr/local/lib/python3.10/dist-packages (from torch->transformers[torch])
(2.20.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.1.105 in

```
/usr/local/lib/python3.10/dist-packages (from torch->transformers[torch])
(12.1.105)
Requirement already satisfied: triton==2.3.0 in /usr/local/lib/python3.10/dist-
packages (from torch->transformers[torch]) (2.3.0)
Requirement already satisfied: nvidia-nvjitlink-cu12 in
/usr/local/lib/python3.10/dist-packages (from nvidia-cusolver-
cu12==11.4.5.107->torch->transformers[torch]) (12.5.40)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers[torch])
(3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests->transformers[torch]) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers[torch])
(2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers[torch])
(2024.6.2)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from
jinja2->torch->transformers[torch]) (2.1.5)
Requirement already satisfied: mpmath<1.4.0,>=1.1.0 in
/usr/local/lib/python3.10/dist-packages (from sympy->torch->transformers[torch])
(1.3.0)
Installing collected packages: accelerate
Successfully installed accelerate-0.31.0
```

```python
import numpy as np
import pandas as pd
import torch
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, f1_score
from transformers import BertTokenizer, BertForSequenceClassification, Trainer,
 ↪TrainingArguments
from transformers import DataCollatorWithPadding
from datasets import Dataset, load_metric
```

```python
# Load your dataset
df = pd.read_excel("DATASET_R1.xlsx")
df.head()
```

```
                    ID                                        Description  \
0           2752994783    // Kirby// they/them // 20 // ENG: ok // FR: ok
1             21571977  Follow us for the latest headlines across the …
2           3382041618                      Growing minds… Changing lives.
3            348478568                                                 NaN
```

```
4   724097168152367104                 One does not simply un-FeelTheBern!


    #Tweets  Date_time_creation_account Language  #Followers  #Friends  \
0      2822   2014 , Aug , 28 , 01:49:22       en          34        68
1      3085   2009 , Feb , 22 , 15:32:48       en       21083        31
2        46   2015 , Aug , 29 , 13:45:01       en          82        68
3      2007   2011 , Aug , 04 , 14:08:52       en          26       270
4     13050   2016 , Apr , 24 , 04:46:32       en         166       287


                           Date&Time  #Favorite  #Retweet  \
0  ['2017', 'Dec', '15', '15:42:41']          0       250
1  ['2017', 'Dec', '15', '15:37:51']          0         0
2  ['2017', 'Dec', '15', '15:37:37']          0         0
3  ['2017', 'Dec', '15', '15:37:16']          0         0
4  ['2017', 'Dec', '15', '15:35:34']          0       151


   Another Tweet Inside  Source            Tweet ID           Retweet ID  \
0                     0  iPhone  941695012122910720  940977667104477185
1                     0  iPhone  941693798056280064       Not a Retweet
2                     0  iPhone  941693735678758912       Not a Retweet
3                     0  Client  941693648634372097       Not a Retweet
4                     0  iPhone  941693221704536064  941044904742006786


        Quote ID  Reply ID Frequency of tweet occurrences State of Tweet
0  Not a quote        NaN                           213.0              r
1  Not a quote        NaN                               1              n
2  Not a quote        NaN                               1              n
3  Not a quote        NaN                               1              n
4  Not a quote        NaN                           140.0              a
```

```python
# Check for null values and remove them
df.dropna(inplace=True)

# Convert DataFrame to HuggingFace Dataset
dataset = Dataset.from_pandas(df)
```

```python
df.isnull().sum()
```

```
ID                            0
Description                   0
#Tweets                       0
Date_time_creation_account    0
Language                      0
#Followers                    0
#Friends                      0
Date&Time                     0
#Favorite                     0
```

```
#Retweet                        0
Another Tweet Inside            0
Source                          0
Tweet ID                        0
Retweet ID                      0
Quote ID                        0
Reply ID                        0
Frequency of tweet occurrences  0
State of Tweet                  0
dtype: int64
```

```python
# Encode labels
label_encoder = LabelEncoder()
df['label'] = label_encoder.fit_transform(df['State of Tweet'])

# Split the dataset
train_texts, val_texts, train_labels, val_labels = train_test_split(
    df['Description'].tolist(), df['label'].tolist(), test_size=0.2,
  ↪random_state=42
)

# Load BERT tokenizer
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')

# Tokenize the texts
train_encodings = tokenizer(train_texts, truncation=True, padding=True)
val_encodings = tokenizer(val_texts, truncation=True, padding=True)

# Convert to datasets
train_dataset = Dataset.from_dict({'input_ids': train_encodings['input_ids'],
  ↪'attention_mask': train_encodings['attention_mask'], 'labels': train_labels})
val_dataset = Dataset.from_dict({'input_ids': val_encodings['input_ids'],
  ↪'attention_mask': val_encodings['attention_mask'], 'labels': val_labels})
```

```
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:89:
UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab
(https://huggingface.co/settings/tokens), set it as secret in your Google Colab
and restart your session.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access
public models or datasets.
  warnings.warn(
```

```
tokenizer_config.json:   0%|          | 0.00/48.0 [00:00<?, ?B/s]

vocab.txt:   0%|          | 0.00/232k [00:00<?, ?B/s]
```

```
tokenizer.json:    0%|              | 0.00/466k [00:00<?, ?B/s]
```

/usr/local/lib/python3.10/dist-packages/huggingface_hub/file_download.py:1132:
FutureWarning: `resume_download` is deprecated and will be removed in version
1.0.0. Downloads always resume when possible. If you want to force a new
download, use `force_download=True`.
  warnings.warn(

```
config.json:    0%|              | 0.00/570 [00:00<?, ?B/s]
```

```python
[ ]: model = BertForSequenceClassification.from_pretrained('bert-base-uncased',␣
     ↪num_labels=len(label_encoder.classes_))
```

```
model.safetensors:    0%|              | 0.00/440M [00:00<?, ?B/s]
```

Some weights of BertForSequenceClassification were not initialized from the
model checkpoint at bert-base-uncased and are newly initialized:
['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.

```python
[ ]: from transformers import TrainingArguments, Trainer

     training_args = TrainingArguments(
         output_dir='./results',            # output directory
         eval_strategy="epoch",             # evaluate each `epoch`
         per_device_train_batch_size=16,    # batch size for training
         per_device_eval_batch_size=16,     # batch size for evaluation
         num_train_epochs=3,                # number of training epochs
         weight_decay=0.01,                 # strength of weight decay
         logging_dir='./logs',              # directory for storing logs
         logging_steps=10,
     )

     # Data collator for dynamic padding
     from transformers import DataCollatorWithPadding

     data_collator = DataCollatorWithPadding(tokenizer=tokenizer)

     # Load metric for evaluation
     from datasets import load_metric

     metric = load_metric("accuracy")

     def compute_metrics(eval_pred):
         logits, labels = eval_pred
         predictions = np.argmax(logits, axis=-1)
         accuracy = accuracy_score(labels, predictions)
         f1 = f1_score(labels, predictions, average='weighted')
```

```
        return {'accuracy': accuracy, 'f1': f1}

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=val_dataset,
    data_collator=data_collator,
    tokenizer=tokenizer,
    compute_metrics=compute_metrics,
)
```

<ipython-input-9-3bbb87e5f3fa>:22: FutureWarning: load_metric is deprecated and will be removed in the next major version of datasets. Use 'evaluate.load' instead, from the new library Evaluate: https://huggingface.co/docs/evaluate
  metric = load_metric("accuracy")
/usr/local/lib/python3.10/dist-packages/datasets/load.py:759: FutureWarning: The repository for accuracy contains custom code which must be executed to correctly load the metric. You can inspect the repository content at https://raw.githubusercontent.com/huggingface/datasets/2.19.2/metrics/accuracy/accuracy.py
You can avoid this message in future by passing the argument `trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this metric from the next major release of `datasets`.
  warnings.warn(

Downloading builder script: 0%| | 0.00/1.65k [00:00<?, ?B/s]

[ ]: trainer.train()

<IPython.core.display.HTML object>

[ ]: TrainOutput(global_step=66, training_loss=0.4927812959208633,
    metrics={'train_runtime': 18.3982, 'train_samples_per_second': 56.256,
    'train_steps_per_second': 3.587, 'total_flos': 38295679556160.0, 'train_loss':
    0.4927812959208633, 'epoch': 3.0})

[ ]: results = trainer.evaluate()
     print(results)

<IPython.core.display.HTML object>

{'eval_loss': 0.4610714614391327, 'eval_accuracy': 0.9080459770114943,
'eval_f1': 0.8725219271897152, 'eval_runtime': 0.3754,
'eval_samples_per_second': 231.758, 'eval_steps_per_second': 15.983, 'epoch':
3.0}

```python
# Predict on validation set
predictions = trainer.predict(val_dataset)

# Get predicted labels
preds = np.argmax(predictions.predictions, axis=-1)

# Calculate accuracy and F1 score
accuracy = accuracy_score(val_labels, preds)
f1 = f1_score(val_labels, preds, average='weighted')

print(f'Validation Accuracy: {accuracy}')
print(f'Validation F1 Score: {f1}')
```
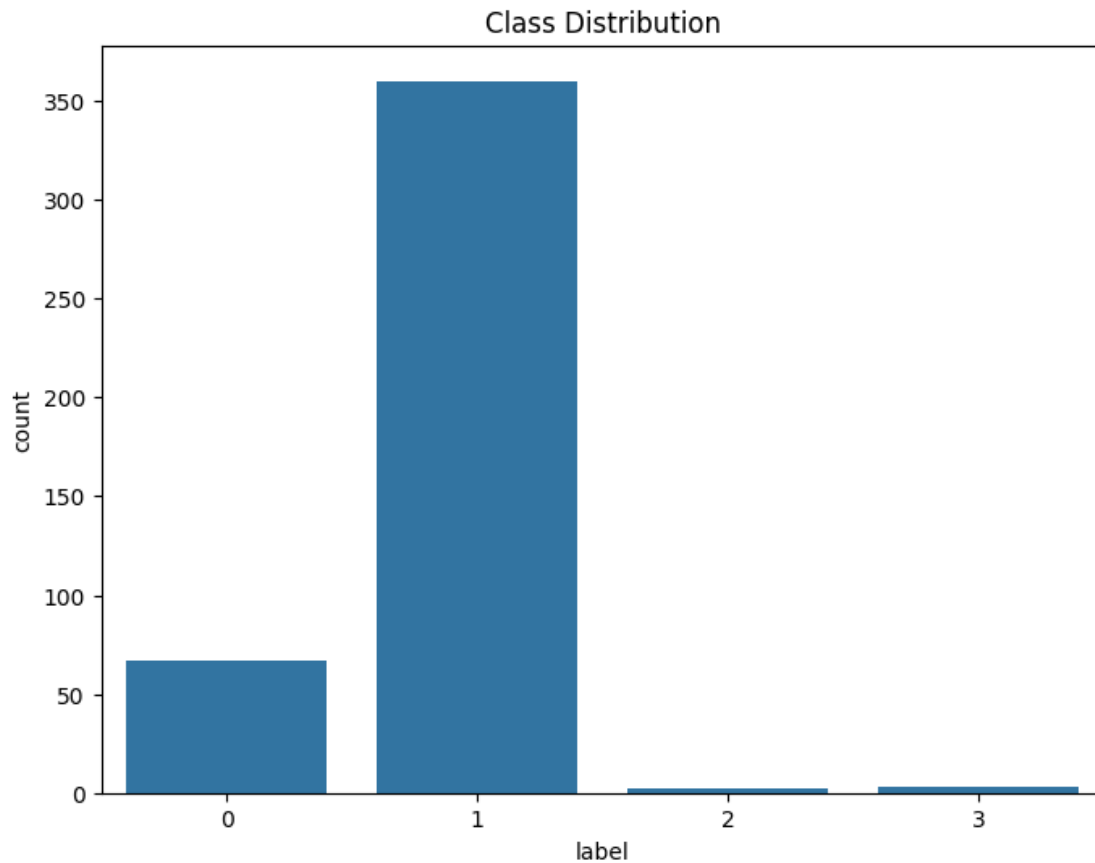
```
<IPython.core.display.HTML object>

Validation Accuracy: 0.9080459770114943
Validation F1 Score: 0.8725219271897152
```

```python
model.save_pretrained('bert_model')
tokenizer.save_pretrained('bert_tokenizer')
```

```
('bert_tokenizer/tokenizer_config.json',
 'bert_tokenizer/special_tokens_map.json',
 'bert_tokenizer/vocab.txt',
 'bert_tokenizer/added_tokens.json')
```

```python
import matplotlib.pyplot as plt
import seaborn as sns
```

```python
# Check class distribution
plt.figure(figsize=(8, 6))
sns.countplot(x='label', data=df)
plt.title('Class Distribution')
plt.show()
```

Class Distribution


```python
# Get training logs
train_logs = trainer.state.log_history

# Initialize lists for metrics
epochs = []
train_loss = []
eval_loss = []
eval_accuracy = []
eval_f1 = []

# Extract the relevant metrics for plotting
for log in train_logs:
    if 'epoch' in log:
        epochs.append(log['epoch'])
    if 'loss' in log:
        train_loss.append(log['loss'])
    if 'eval_loss' in log:
        eval_loss.append(log['eval_loss'])
    if 'eval_accuracy' in log:
```

```
        eval_accuracy.append(log['eval_accuracy'])
    if 'eval_f1' in log:
        eval_f1.append(log['eval_f1'])

# Ensure all lists are the same length
min_len = min(len(epochs), len(train_loss), len(eval_loss), len(eval_accuracy),␣
 ↪len(eval_f1))
epochs = epochs[:min_len]
train_loss = train_loss[:min_len]
eval_loss = eval_loss[:min_len]
eval_accuracy = eval_accuracy[:min_len]
eval_f1 = eval_f1[:min_len]
```
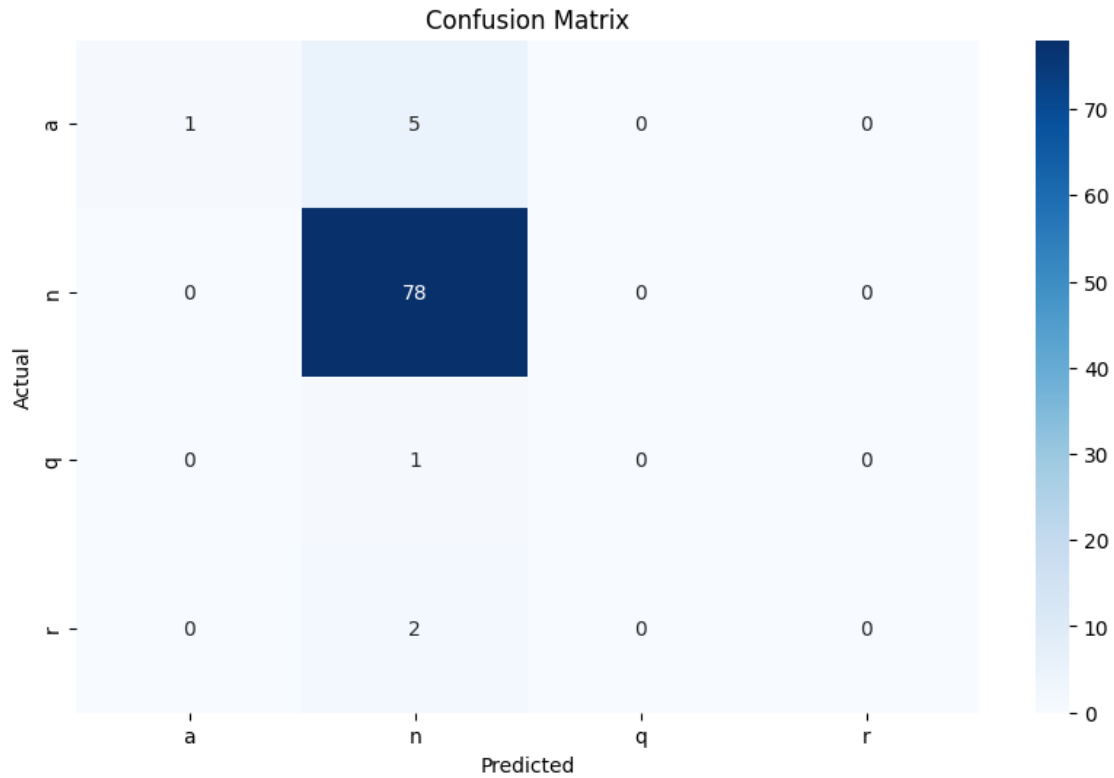
```
[ ]: from sklearn.metrics import accuracy_score, f1_score, confusion_matrix
```
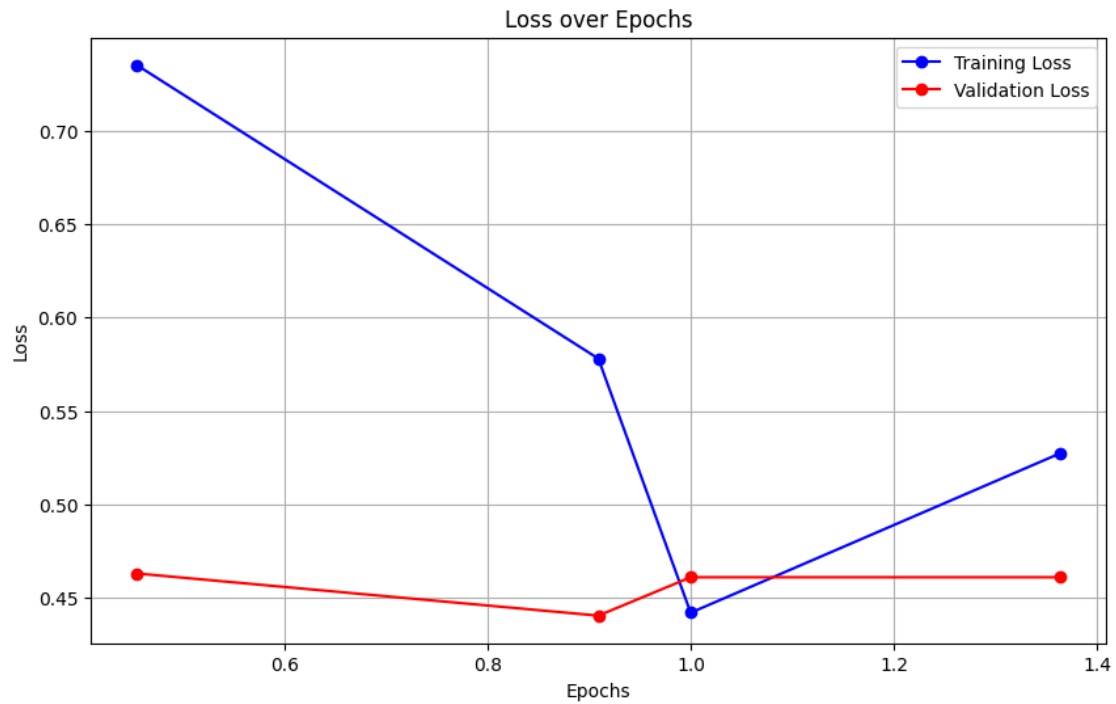
```
[ ]: # Confusion matrix
     y_pred = trainer.predict(val_dataset).predictions.argmax(axis=-1)
     conf_matrix = confusion_matrix(val_labels, y_pred)
     plt.figure(figsize=(10, 6))
     sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues',␣
      ↪xticklabels=label_encoder.classes_, yticklabels=label_encoder.classes_)
     plt.title('Confusion Matrix')
     plt.xlabel('Predicted')
     plt.ylabel('Actual')
     plt.savefig('confusion_matrix.png')
     plt.show()
```
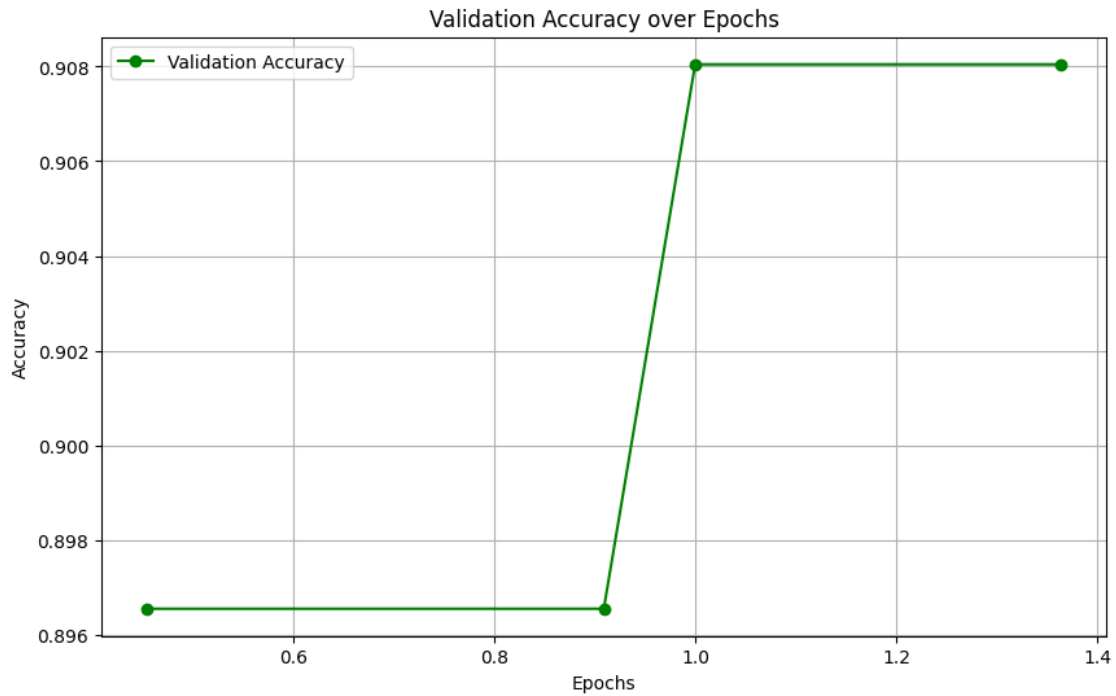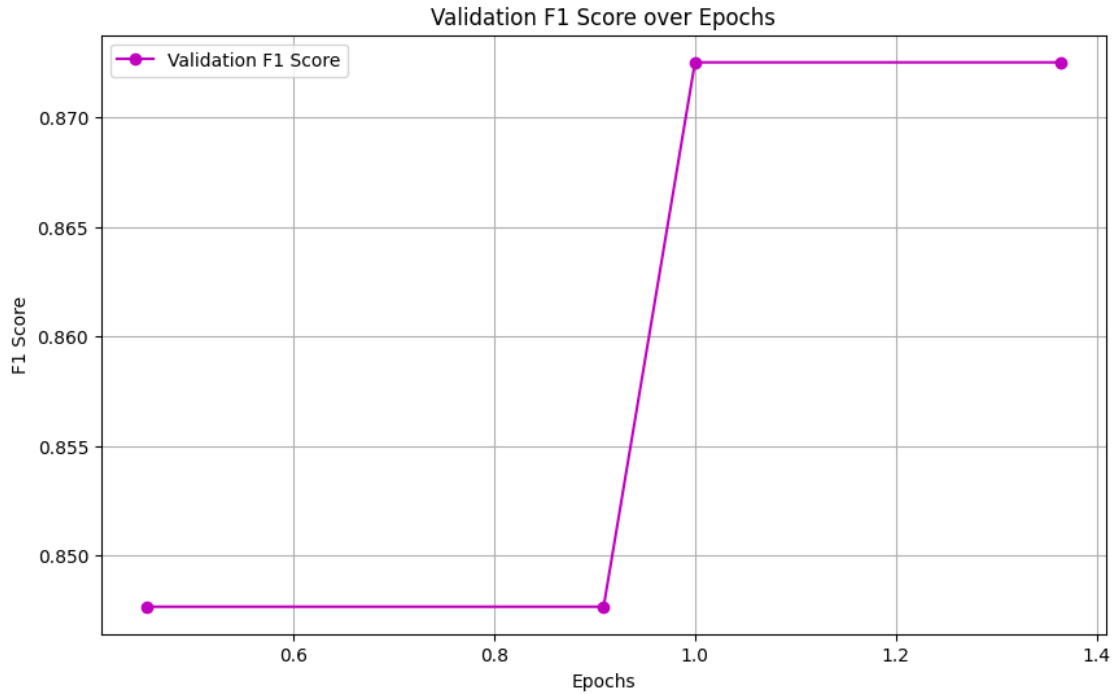
```
<IPython.core.display.HTML object>
```

Confusion Matrix

Actual \ Predicted	a	n	q	r
a	1	5	0	0
n	0	78	0	0
q	0	1	0	0
r	0	2	0	0

```python
# Plot loss over epochs
plt.figure(figsize=(10, 6))
plt.plot(epochs, train_loss, marker='o', linestyle='-', color='b',
    label='Training Loss')
plt.plot(epochs, eval_loss, marker='o', linestyle='-', color='r',
    label='Validation Loss')
plt.title('Loss over Epochs')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
plt.grid(True)
plt.savefig('loss_plot.png')
plt.show()
```

Loss over Epochs

```python
# Plot accuracy over epochs
plt.figure(figsize=(10, 6))
plt.plot(epochs, eval_accuracy, marker='o', linestyle='-', color='g',␣
 ↪label='Validation Accuracy')
plt.title('Validation Accuracy over Epochs')
plt.xlabel('Epochs')
plt.ylabel('Accuracy')
plt.legend()
plt.grid(True)
plt.savefig('accuracy_plot.png')
plt.show()
```

Validation Accuracy over Epochs

```python
# Plot F1 score over epochs
plt.figure(figsize=(10, 6))
plt.plot(epochs, eval_f1, marker='o', linestyle='-', color='m',␣
 ↪label='Validation F1 Score')
plt.title('Validation F1 Score over Epochs')
plt.xlabel('Epochs')
plt.ylabel('F1 Score')
plt.legend()
plt.grid(True)
plt.savefig('f1_score_plot.png')
plt.show()
```

Validation F1 Score over Epochs

[ ]:

[ ]:

[ ]: `!pip install streamlit`

```
Collecting streamlit
  Downloading streamlit-1.35.0-py2.py3-none-any.whl (8.6 MB)
                          8.6/8.6 MB
22.4 MB/s eta 0:00:00
Requirement already satisfied: altair<6,>=4.0 in
/usr/local/lib/python3.10/dist-packages (from streamlit) (4.2.2)
Requirement already satisfied: blinker<2,>=1.0.0 in /usr/lib/python3/dist-
packages (from streamlit) (1.4)
Requirement already satisfied: cachetools<6,>=4.0 in
/usr/local/lib/python3.10/dist-packages (from streamlit) (5.3.3)
Requirement already satisfied: click<9,>=7.0 in /usr/local/lib/python3.10/dist-
packages (from streamlit) (8.1.7)
Requirement already satisfied: numpy<2,>=1.19.3 in
/usr/local/lib/python3.10/dist-packages (from streamlit) (1.25.2)
Requirement already satisfied: packaging<25,>=16.8 in
/usr/local/lib/python3.10/dist-packages (from streamlit) (24.0)
Requirement already satisfied: pandas<3,>=1.3.0 in
/usr/local/lib/python3.10/dist-packages (from streamlit) (2.0.3)
Requirement already satisfied: pillow<11,>=7.1.0 in
```

/usr/local/lib/python3.10/dist-packages (from streamlit) (9.4.0)
Requirement already satisfied: protobuf<5,>=3.20 in
/usr/local/lib/python3.10/dist-packages (from streamlit) (3.20.3)
Requirement already satisfied: pyarrow>=7.0 in /usr/local/lib/python3.10/dist-
packages (from streamlit) (14.0.2)
Requirement already satisfied: requests<3,>=2.27 in
/usr/local/lib/python3.10/dist-packages (from streamlit) (2.32.3)
Requirement already satisfied: rich<14,>=10.14.0 in
/usr/local/lib/python3.10/dist-packages (from streamlit) (13.7.1)
Requirement already satisfied: tenacity<9,>=8.1.0 in
/usr/local/lib/python3.10/dist-packages (from streamlit) (8.3.0)
Requirement already satisfied: toml<2,>=0.10.1 in
/usr/local/lib/python3.10/dist-packages (from streamlit) (0.10.2)
Requirement already satisfied: typing-extensions<5,>=4.3.0 in
/usr/local/lib/python3.10/dist-packages (from streamlit) (4.12.1)
Collecting gitpython!=3.1.19,<4,>=3.0.7 (from streamlit)
  Downloading GitPython-3.1.43-py3-none-any.whl (207 kB)
                            207.3/207.3

kB 21.6 MB/s eta 0:00:00
Collecting pydeck<1,>=0.8.0b4 (from streamlit)
  Downloading pydeck-0.9.1-py2.py3-none-any.whl (6.9 MB)
                            6.9/6.9 MB
57.0 MB/s eta 0:00:00
Requirement already satisfied: tornado<7,>=6.0.3 in
/usr/local/lib/python3.10/dist-packages (from streamlit) (6.3.3)
Collecting watchdog>=2.1.5 (from streamlit)
  Downloading watchdog-4.0.1-py3-none-manylinux2014_x86_64.whl (83 kB)
                            83.0/83.0 kB
9.9 MB/s eta 0:00:00
Requirement already satisfied: entrypoints in
/usr/local/lib/python3.10/dist-packages (from altair<6,>=4.0->streamlit) (0.4)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages
(from altair<6,>=4.0->streamlit) (3.1.4)
Requirement already satisfied: jsonschema>=3.0 in
/usr/local/lib/python3.10/dist-packages (from altair<6,>=4.0->streamlit)
(4.19.2)
Requirement already satisfied: toolz in /usr/local/lib/python3.10/dist-packages
(from altair<6,>=4.0->streamlit) (0.12.1)
Collecting gitdb<5,>=4.0.1 (from gitpython!=3.1.19,<4,>=3.0.7->streamlit)
  Downloading gitdb-4.0.11-py3-none-any.whl (62 kB)
                            62.7/62.7 kB
7.3 MB/s eta 0:00:00
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.10/dist-packages (from pandas<3,>=1.3.0->streamlit)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas<3,>=1.3.0->streamlit) (2023.4)

```
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-
packages (from pandas<3,>=1.3.0->streamlit) (2024.1)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests<3,>=2.27->streamlit)
(3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests<3,>=2.27->streamlit) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests<3,>=2.27->streamlit)
(2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests<3,>=2.27->streamlit)
(2024.6.2)
Requirement already satisfied: markdown-it-py>=2.2.0 in
/usr/local/lib/python3.10/dist-packages (from rich<14,>=10.14.0->streamlit)
(3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in
/usr/local/lib/python3.10/dist-packages (from rich<14,>=10.14.0->streamlit)
(2.16.1)
Collecting smmap<6,>=3.0.1 (from
gitdb<5,>=4.0.1->gitpython!=3.1.19,<4,>=3.0.7->streamlit)
  Downloading smmap-5.0.1-py3-none-any.whl (24 kB)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from jinja2->altair<6,>=4.0->streamlit)
(2.1.5)
Requirement already satisfied: attrs>=22.2.0 in /usr/local/lib/python3.10/dist-
packages (from jsonschema>=3.0->altair<6,>=4.0->streamlit) (23.2.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
/usr/local/lib/python3.10/dist-packages (from
jsonschema>=3.0->altair<6,>=4.0->streamlit) (2023.12.1)
Requirement already satisfied: referencing>=0.28.4 in
/usr/local/lib/python3.10/dist-packages (from
jsonschema>=3.0->altair<6,>=4.0->streamlit) (0.35.1)
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.10/dist-
packages (from jsonschema>=3.0->altair<6,>=4.0->streamlit) (0.18.1)
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.10/dist-
packages (from markdown-it-py>=2.2.0->rich<14,>=10.14.0->streamlit) (0.1.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.8.2->pandas<3,>=1.3.0->streamlit) (1.16.0)
Installing collected packages: watchdog, smmap, pydeck, gitdb, gitpython,
streamlit
Successfully installed gitdb-4.0.11 gitpython-3.1.43 pydeck-0.9.1 smmap-5.0.1
streamlit-1.35.0 watchdog-4.0.1
```

```python
with open('app.py', 'w') as f:
    f.write("""
import streamlit as st
```

```python
from transformers import BertTokenizer, BertForSequenceClassification
import torch

@st.cache(allow_output_mutation=True)
def load_model():
    model = BertForSequenceClassification.from_pretrained('bert_model')
    tokenizer = BertTokenizer.from_pretrained('bert_tokenizer')
    return model, tokenizer

model, tokenizer = load_model()

st.title("Tweet State Classification")
user_input = st.text_area("Enter the tweet description")

if st.button("Classify"):
    inputs = tokenizer(user_input, return_tensors="pt", truncation=True,
 padding=True)
    outputs = model(**inputs)
    predictions = torch.argmax(outputs.logits, dim=-1).item()
    st.write(f"Predicted State: {predictions}")
""")
```

```
[ ]: ! wget -q -O - ipv4.icanhazip.com
```

35.233.138.80

```
[ ]: ! streamlit run app.py & npx localtunnel --port 8501
```

Collecting usage statistics. To deactivate, set browser.gatherUsageStats to
false.

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://172.28.0.12:8501
  External URL: http://35.233.138.80:8501

npx: installed 22 in 3.809s
your url is: https://swift-seals-hear.loca.lt
2024-06-11 18:15:16.316 `st.cache` is deprecated. Please use one of Streamlit's
new caching commands,
`st.cache_data` or `st.cache_resource`.

More information [in our docs](https://docs.streamlit.io/library/advanced-
features/caching).

21

2024-06-11 18:16:17.419 `st.cache` is deprecated. Please use one of Streamlit's new caching commands,
`st.cache_data` or `st.cache_resource`.

More information [in our docs](https://docs.streamlit.io/library/advanced-features/caching).
  Stopping…
^C