```
from google.colab import files

uploaded = files.upload()
```

Choose Files updated_da…bels (1).csv

- **updated_dataset_with_dialect_labels (1).csv**(text/csv) - 761664 bytes, last modified: 12/9/2024 - 100% done

```
Saving updated_dataset_with_dialect_labels (1).csv to updated_dataset_with_dialect_labels (1).csv
```

```
print(uploaded.keys())
```

```
dict_keys(['updated_dataset_with_dialect_labels (1).csv'])
```

```
import pandas as pd
from transformers import AutoTokenizer
import torch
from collections import Counter
import re

dataset = pd.read_csv("updated_dataset_with_dialect_labels (1).csv")

label_map = {"india": 0, "usa": 1, "united kingdom": 2}

def clean_label(label):
    label = label.lower()
    label = re.sub(r"[^a-zA-Z\s]", "", label)
    label = label.strip()
    return label

original_labels = dataset["Region"].tolist()
cleaned_labels = [clean_label(label) for label in original_labels]

labels = [label_map.get(label, -1) for label in cleaned_labels]

print("Label counts before mapping:", Counter(original_labels))
print("Label counts after cleaning:", Counter(cleaned_labels))
print("Label counts after mapping:", Counter(labels))

texts = dataset["Post_and_Comments"].tolist()

tokenizer = AutoTokenizer.from_pretrained("bert-base-uncased")

encoded_data = tokenizer(texts, padding=True, truncation=True, return_tensors="pt")

torch.save(encoded_data, 'tokenized_output.pt')

print("Sample tokenized input IDs:", encoded_data['input_ids'][:3])
print("Sample attention masks:", encoded_data['attention_mask'][:3])

print("Tokenization complete. Tokenized data saved to 'tokenized_output.pt'.")
```

```
Label counts before mapping: Counter({'United Kingdom': 553, 'USA': 511, 'India': 264})
Label counts after cleaning: Counter({'united kingdom': 553, 'usa': 511, 'india': 264})
Label counts after mapping: Counter({2: 553, 1: 511, 0: 264})
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_auth.py:94: 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 se
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: 100%          48.0/48.0 [00:00<00:00, 1.66kB/s]

config.json: 100%          570/570 [00:00<00:00, 23.4kB/s]

vocab.txt: 100%          232k/232k [00:00<00:00, 1.77MB/s]

tokenizer.json: 100%          466k/466k [00:00<00:00, 3.37MB/s]
```

```
Sample tokenized input IDs: tensor([[  101,  5976, 16510, 15512,  8509, 17763, 11139, 28919,  7662,  3211,
          2011, 18155,  2140,  5564, 27612,  2362,  2561,  2474, 10023,  7206,
         14383,  3981,  2102,  7486,  7093, 27612,  2501, 16510,  3469,  7017,
          2685,  4012, 19362,  3320, 14592,  5119, 16510,  5973,  7017,  2685,
          2048, 27612,  2501,  5717, 15512,  7486,  2862,  9065,  3616,  3320,
         14592, 27612,  2501,  3867,  3867, 15512,  4414,  2387,  3867,  3867,
          8554,  4414, 13660, 11200, 14592, 16913,  2072, 15030,  8081, 15030,
          3972,  3388,  1054,  2860,  1043,  4859,  2271,  2994,  2185,  2695,
          3904,  7615,  5294, 14303, 16341,  9061, 18155,  4877,  2695,  2092,
         16770,  2860,  2860, 13088, 22367,  4183,  9006,  6657,  9032, 14540,
          2546,  8950,  6292, 16257,  2497,  4160,  2174,  2695,  2288, 16908,
          4215,  2695,  2288, 16476,  2593,  2111,  2123,  2102,  2215,  2156,
          7072,  2579,  4536,  2560,  3191,  3720,  2831, 14303, 15512, 21040,
         11703, 16416,  2015,  5152,  3484, 27612,   102,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0],
        [  101,  3198,  2634, 11689,  6160, 15544, 16089,  3022,  3198,  2634,
         11689, 10861,  3089,  2166,  2634,  2166,  6505, 11689,  3531,  2562,
          2568,  3582, 21766,  2140,  2327,  2504,  7928, 24501,  2121,  2615,
         10861,  3089,  2576,  8466,  3276, 10861,  3089,  7141, 25269, 10581,
          9285,  5605, 22254,  2401,  3531,  3046,  3945,  4274,  3198,  2393,
          2070,  3775,  2213,  3437,  4274,  3945,  2185,  3080,   100,  4658,
         15239,  2015,  2634,  4497, 10047,  2431,  2796,  3942,  2034,  2051,
          3204, 10047,  2036,  4827,  2866,  4658,  2611,  7340,  2052, 20228,
          6305,  2066,  7367, 13471,  2638, 13708, 17278,  2239,  5506,  4063,
          4385,  4687,  3087,  2071, 16755,  5662,  2015,  2634, 10047,  4919,
          2298,  2796, 11216, 23544,  2015,  4638,  4067,  4931,  5665,  3604,
          2634, 27249,  2285, 21981,  2094,  5665,  5463,  2215,  2202,  8412,
         12098,  4921, 10014,  5956,  3229, 10050,  6097,  8654,  4013,  2151,
          2705,  4737, 13586,  2123,  2102,  2215,  3941,  4638,  2378,  1048,
          3726,  4950,  5477,  8490,  2627,  2123,  2102,   102,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
             0],
        [  101,  9587,  3511,  2213, 16950, 26068,  2099, 25869,  2290,  2203,
         25121, 12601,  8499, 11109,  2634,  7173,  2265,  2678,  5223, 12256,
```