Topic Modelling and Text Categorisation

```
!pwd
```

```
/content
```

```
!pip install -U datasets
!pip install -U simpletransformers
```

```
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->simpletransformer
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn->simpletransf
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn->simpl
Requirement already satisfied: altair<6,>=4.0 in /usr/local/lib/python3.11/dist-packages (from streamlit->simpletransfor
Requirement already satisfied: blinker<2,>=1.5.0 in /usr/local/lib/python3.11/dist-packages (from streamlit->simpletrans
Requirement already satisfied: cachetools<6,>=4.0 in /usr/local/lib/python3.11/dist-packages (from streamlit->simpletran
Requirement already satisfied: pillow<12,>=7.1.0 in /usr/local/lib/python3.11/dist-packages (from streamlit->simpletrans
Requirement already satisfied: tenacity<10,>=8.1.0 in /usr/local/lib/python3.11/dist-packages (from streamlit->simpletra
Requirement already satisfied: toml<2,>=0.10.1 in /usr/local/lib/python3.11/dist-packages (from streamlit->simpletransfo
Collecting watchdog<7,>=2.1.5 (from streamlit->simpletransformers)
  Downloading watchdog-6.0.0-py3-none-manylinux2014_x86_64.whl.metadata (44 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 44.3/44.3 kB 3.7 MB/s eta 0:00:00
Collecting pydeck<1,>=0.8.0b4 (from streamlit->simpletransformers)
  Downloading pydeck-0.9.1-py2.py3-none-any.whl.metadata (4.1 kB)
Requirement already satisfied: tornado<7,>=6.0.3 in /usr/local/lib/python3.11/dist-packages (from streamlit->simpletrans
Requirement already satisfied: absl-py>=0.4 in /usr/local/lib/python3.11/dist-packages (from tensorboard->simpletransfor
Requirement already satisfied: grpcio>=1.48.2 in /usr/local/lib/python3.11/dist-packages (from tensorboard->simpletransf
Requirement already satisfied: markdown>=2.6.8 in /usr/local/lib/python3.11/dist-packages (from tensorboard->simpletrans
Requirement already satisfied: six>1.9 in /usr/local/lib/python3.11/dist-packages (from tensorboard->simpletransformers)
Requirement already satisfied: tensorboard-data-server<0.8.0,>=0.7.0 in /usr/local/lib/python3.11/dist-packages (from te
Requirement already satisfied: werkzeug>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from tensorboard->simpletrans
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from altair<6,>=4.0->streamlit->simple
Requirement already satisfied: jsonschema>=3.0 in /usr/local/lib/python3.11/dist-packages (from altair<6,>=4.0->streamli
Requirement already satisfied: narwhals>=1.14.2 in /usr/local/lib/python3.11/dist-packages (from altair<6,>=4.0->streaml
Requirement already satisfied: aiohttp!=4.0.0a0,!=4.0.0a1 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]<
Requirement already satisfied: gitdb<5,>=4.0.1 in /usr/local/lib/python3.11/dist-packages (from gitpython!=3.1.29,>=1.0.
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic<3->wandb
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic<3->wandb>
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic<3->wan
Requirement already satisfied: MarkupSafe>=2.1.1 in /usr/local/lib/python3.11/dist-packages (from werkzeug>=1.0.1->tenso
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0
Requirement already satisfied: smmap<6,>=3.0.1 in /usr/local/lib/python3.11/dist-packages (from gitdb<5,>=4.0.1->gitpyth
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /usr/local/lib/python3.11/dist-packages (from jso
Requirement already satisfied: referencing>=0.28.4 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=3.0->alt
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=3.0->altair<6
Downloading simpletransformers-0.70.1-py3-none-any.whl (316 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 316.3/316.3 kB 19.7 MB/s eta 0:00:00
Downloading streamlit-1.45.1-py3-none-any.whl (9.9 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 9.9/9.9 MB 82.9 MB/s eta 0:00:00
Downloading tensorboardX-2.6.2.2-py2.py3-none-any.whl (101 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 101.7/101.7 kB 8.4 MB/s eta 0:00:00
Downloading pydeck-0.9.1-py2.py3-none-any.whl (6.9 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.9/6.9 MB 99.2 MB/s eta 0:00:00
Downloading watchdog-6.0.0-py3-none-manylinux2014_x86_64.whl (79 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 79.1/79.1 kB 7.1 MB/s eta 0:00:00
Building wheels for collected packages: seqeval
  Building wheel for seqeval (setup.py) ... done
  Created wheel for seqeval: filename=seqeval-1.2.2-py3-none-any.whl size=16162 sha256=f5b13bd34c53e92b19fa04da6dc97f924
  Stored in directory: /root/.cache/pip/wheels/bc/92/f0/243288f899c2eacdfa8c5f9aede4c71a9bad0ee26a01dc5ead
Successfully built seqeval
Installing collected packages: watchdog, tensorboardx, pydeck, seqeval, streamlit, simpletransformers
Successfully installed pydeck-0.9.1 seqeval-1.2.2 simpletransformers-0.70.1 streamlit-1.45.1 tensorboardx-2.6.2.2 watchd
```

```
from datasets import load_dataset
import pandas as pd

ds = load_dataset("AmazonScience/mintaka", 'en')

train_df = pd.concat([pd.DataFrame(ds['train']), pd.DataFrame(ds['validation'])])[['question', 'category']]
test_df = pd.DataFrame(ds['test'])[['question', 'category']]
train_df.dropna(inplace=True)
test_df.dropna(inplace=True)
```

```
/usr/local/lib/python3.11/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),
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(
```

```
README.md: 100%                                   7.47k/7.47k [00:00<00:00, 155kB/s]
mintaka.py: 100%                                7.19k/7.19k [00:00<00:00, 159kB/s]
0000.parquet: 100%                                1.45M/1.45M [00:00<00:00, 13.9MB/s]
0000.parquet: 100%                                252k/252k [00:00<00:00, 12.3MB/s]
0000.parquet: 100%                                468k/468k [00:00<00:00, 10.9MB/s]
Generating train split: 100%                          14000/14000 [00:00<00:00, 115303.93 examples/s]
Generating validation split: 100%                        2000/2000 [00:00<00:00, 41796.33 examples/s]
Generating test split: 100%                          4000/4000 [00:00<00:00, 74926.38 examples/s]
```

```python
possible_topics_train = train_df['category'].unique()
possible_topics_test = test_df['category'].unique()
possible_topics = list(set(possible_topics_train) | set(possible_topics_test))
print(possible_topics)
print(len(possible_topics))
```

```
['movies', 'music', 'history', 'politics', 'sports', 'videogames', 'books', 'geography']
8
```

```python
print(train_df.head)
print(train_df.shape)
print(test_df.head)
print(test_df.shape)
```

```python
from simpletransformers.classification import ClassificationModel, ClassificationArgs

model_args = ClassificationArgs()

model_args.overwrite_output_dir=True
model_args.evaluate_during_training=True
model_args.no_save = True

model_args.max_seq_length=256
model_args.use_early_stopping=True
model_args.early_stopping_delta=0.01
model_args.early_stopping_metric='eval_loss'
model_args.early_stopping_metric_minimize=True
model_args.early_stopping_patience=2
model_args.evaluate_during_training_steps=32
model_args.reprocess_input_data=True
model_args.manual_seed=4
model_args.use_multiprocessing=True
model_args.labels_list=possible_topics
model_args.wandb_project="Topic-Catagorization-Sweep"
```

```python
import wandb

def train_model():
  wandb.init()
  model = ClassificationModel("distilbert", "distilbert-base-uncased", num_labels=len(possible_topics), args=model_args, use
  model.train_model(train_df, eval_df=test_df)
```

```python
wandb.login()

sweep_configuration = {
    "method": "grid",
    "metric": {"goal": "minimize", "name": "eval_loss"},
    "parameters": {
        "train_epochs": {"values": [8, 10, 12, 16]},
        "train_batch_size": {"values": [16, 32, 64]},
        "learning_rate": {"values": [1e-6, 5e-5, 1e-5, 5e-4]},
    }
}
sweep_id = input("Please enter an existing sweep id if you want to continue an existing sweep (leave blank for a new sweep):
sweep_id = None if sweep_id == "" else sweep_id

if sweep_id is None:
    sweep_id = wandb.sweep(sweep=sweep_configuration, project="Topic-Catagorization-Sweep")
```

```
wandb.agent(sweep_id, function=train_model, project='Topic-Catagorization-Sweep')
```

```
eval_df = pd.read_csv('./sentiment-topic-test.tsv', sep='\t')
eval_df = eval_df.dropna()
eval_df.drop(['sentence_id', 'sentiment'], axis=1, inplace=True)
eval_df.columns = ['question', 'category']
eval_df['category'] = eval_df['category'].replace({'book': 'books', 'movie': 'movies'})
eval_categories = eval_df['category'].unique()
possible_topics = list(set(possible_topics_test) | set(eval_categories) | set(possible_topics_train))
```

```
print(possible_topics)
print(len(possible_topics))
```

```
['music', 'movies', 'history', 'politics', 'sports', 'videogames', 'books', 'geography']
8
```

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

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

```
from simpletransformers.classification import ClassificationModel, ClassificationArgs

best_model_args = ClassificationArgs()

best_model_args.overwrite_output_dir=True
best_model_args.evaluate_during_training=True
best_model_args.no_save = True

best_model_args.max_seq_length=256
best_model_args.use_early_stopping=True
best_model_args.early_stopping_delta=0.01
best_model_args.early_stopping_metric='eval_loss'
best_model_args.early_stopping_metric_minimize=True
best_model_args.early_stopping_patience=2
best_model_args.evaluate_during_training_steps=32
best_model_args.reprocess_input_data=True
best_model_args.manual_seed=25
best_model_args.use_multiprocessing=True
best_model_args.labels_list=possible_topics


best_model_args.num_train_epochs=16
best_model_args.train_batch_size=64
best_model_args.learning_rate=5e-5

best_model = ClassificationModel("distilbert", "distilbert-base-uncased", num_labels=len(possible_topics), args=best_model_a
```

```
config.json: 100%                                483/483 [00:00<00:00, 42.5kB/s]
Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. Falling back to regular HTTP download.
WARNING:huggingface_hub.file_download:Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. F
model.safetensors: 100%                                268M/268M [00:01<00:00, 309MB/s]
Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at distilbert-base-un
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
tokenizer_config.json: 100%                                48.0/48.0 [00:00<00:00, 5.13kB/s]
vocab.txt: 100%                                232k/232k [00:00<00:00, 1.43MB/s]
tokenizer.json: 100%                                466k/466k [00:00<00:00, 2.85MB/s]
```

```
best_model.train_model(pd.concat([train_df, test_df]), eval_df=eval_df)
```

```
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:610: UserWarning: Data
  warnings.warn(
100%                                                  40/40 [00:08<00:00,  5.98it/s]

Epoch 2 of 16:   6%                                   1/16 [01:56<23:36, 94.45s/it]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:882: FutureWarning: `t
  scaler = amp.GradScaler()
Epochs 1/16. Running Loss:    0.1123: 100%                        313/313 [01:34<00:00,  3.81it/s]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
     1/0 [00:00<00:00, 13.17it/s]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
     1/0 [00:00<00:00, 12.11it/s]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
     1/0 [00:00<00:00,  8.45it/s]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
     1/0 [00:00<00:00, 12.84it/s]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
     1/0 [00:00<00:00, 12.10it/s]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
     1/0 [00:00<00:00, 12.19it/s]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
     1/0 [00:00<00:00,  8.81it/s]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
     1/0 [00:00<00:00, 12.02it/s]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
     1/0 [00:00<00:00, 12.48it/s]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
     1/0 [00:00<00:00, 10.70it/s]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
Epochs 2/16. Running Loss:    0.1895:  22%                        70/313 [00:22<01:14,  3.26it/s]
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
```

```
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
```

1/0 [00:00<00:00, 10.89it/s]

```
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
```

1/0 [00:00<00:00, 8.71it/s]

```
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:905: FutureWarning: `t
  with amp.autocast():
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1453: UserWarning: Dat
  warnings.warn(
```

1/0 [00:00<00:00, 12.14it/s]

```
/usr/local/lib/python3.11/dist-packages/simpletransformers/classification/classification_model.py:1505: FutureWarning: `
  with amp.autocast():
(384,
 defaultdict(list,
             {'global_step': [32,
               64,
               96,
               128,
               160,
               192,
               224,
               256,
               288,
               313,
               320,
               352,
               384],
              'train_loss': [2.0601959228515625,
               1.7057952880859375,
               0.8503837585449219,
               0.5328502655029297,
               0.2613534927368164,
               0.25365352630615234,
               0.3656768798828125,
               0.3202791213989258,
               0.236405611038208,
               0.11225581169128418,
               0.12090396881103516,
               0.07753157615661621,
               0.18953460454940796],
              'mcc': [0.0,
               np.float64(0.75),
               np.float64(0.7603756252309958),
               np.float64(0.6968731476445666),
               np.float64(0.6968731476445666),
               np.float64(0.7085662394599952),
               np.float64(0.7739527433912545),
               np.float64(0.7739527433912545),
               np.float64(0.8333333333333334),
               np.float64(0.7534965224510292),
               np.float64(0.8418444422200311),
               np.float64(0.6405126152203485),
               np.float64(0.7603756252309958)],
              'eval_loss': [2.095160484313965,
               1.8145616054534912,
               1.1092122793197632,
               0.7959662675857544,
               0.6227790117263794,
               0.6657986044883728,
               0.4731394350528717,
               0.4596574604511261,
               0.36544036865234375,
               0.5016581416130066,
               0.4691687822341919,
               0.706573486328125,
               0.5369746088981628]}))
```