1 Propaganda Dataset Classification Functional Test MVP with BERT base

April 2, 2025

[32]:
```
#@title Install Packages
```

[1]:
```
!pip install -q transformers
!pip install -q torchinfo
!pip install -q datasets
!pip install -q evaluate
```

```
                                        491.2/491.2 kB
7.7 MB/s eta 0:00:00
                                        116.3/116.3 kB
7.4 MB/s eta 0:00:00
                                        183.9/183.9 kB
6.5 MB/s eta 0:00:00
                                        143.5/143.5 kB
4.4 MB/s eta 0:00:00
                                        194.8/194.8 kB
8.1 MB/s eta 0:00:00
```

```
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.
torch 2.6.0+cu124 requires nvidia-cublas-cu12==12.4.5.8; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cublas-cu12
12.5.3.2 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cuda-cupti-cu12==12.4.127; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cuda-cupti-cu12
12.5.82 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cuda-nvrtc-cu12==12.4.127; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cuda-nvrtc-cu12
12.5.82 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cuda-runtime-cu12==12.4.127; platform_system
== "Linux" and platform_machine == "x86_64", but you have nvidia-cuda-runtime-
cu12 12.5.82 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cudnn-cu12==9.1.0.70; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cudnn-cu12
9.3.0.75 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cufft-cu12==11.2.1.3; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cufft-cu12
11.2.3.61 which is incompatible.
torch 2.6.0+cu124 requires nvidia-curand-cu12==10.3.5.147; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-curand-cu12
10.3.6.82 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cusolver-cu12==11.6.1.9; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cusolver-cu12
11.6.3.83 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cusparse-cu12==12.3.1.170; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cusparse-cu12
12.5.1.3 which is incompatible.
torch 2.6.0+cu124 requires nvidia-nvjitlink-cu12==12.4.127; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-nvjitlink-cu12
12.5.82 which is incompatible.
gcsfs 2025.3.0 requires fsspec==2025.3.0, but you have fsspec 2024.12.0 which is
incompatible.
```


84.0/84.0 kB

```
2.4 MB/s eta 0:00:00
```

[21]:
```
!sudo apt-get update
! sudo apt-get install tree
```

```
Get:1 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Get:2 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
[3,632 B]
Hit:3 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64
InRelease
Hit:4 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:5 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Get:6 https://r2u.stat.illinois.edu/ubuntu jammy InRelease [6,555 B]
Get:7 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
Get:8 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ Packages [70.3
kB]
Get:9 http://security.ubuntu.com/ubuntu jammy-security/restricted amd64 Packages
[3,972 kB]
Hit:10 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:11 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy
InRelease
Get:12 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages
[2,773 kB]
Hit:13 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:14 https://r2u.stat.illinois.edu/ubuntu jammy/main all Packages [8,802 kB]
Get:15 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages
[1,540 kB]
Get:16 https://r2u.stat.illinois.edu/ubuntu jammy/main amd64 Packages [2,685 kB]
Get:17 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [3,081
kB]
Fetched 23.3 MB in 3s (8,070 kB/s)
Reading package lists… Done
W: Skipping acquire of configured file 'main/source/Sources' as repository
'https://r2u.stat.illinois.edu/ubuntu jammy InRelease' does not seem to provide
it (sources.list entry misspelt?)
Reading package lists… Done
Building dependency tree… Done
Reading state information… Done
The following NEW packages will be installed:
  tree
0 upgraded, 1 newly installed, 0 to remove and 32 not upgraded.
Need to get 47.9 kB of archives.
After this operation, 116 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tree amd64 2.0.2-1
[47.9 kB]
Fetched 47.9 kB in 0s (350 kB/s)
debconf: unable to initialize frontend: Dialog
```

```
debconf: (No usable dialog-like program is installed, so the dialog based
frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line 78,
<> line 1.)
debconf: falling back to frontend: Readline
debconf: unable to initialize frontend: Readline
debconf: (This frontend requires a controlling tty.)
debconf: falling back to frontend: Teletype
dpkg-preconfigure: unable to re-open stdin:
Selecting previously unselected package tree.
(Reading database … 126210 files and directories currently installed.)
Preparing to unpack …/tree_2.0.2-1_amd64.deb …
Unpacking tree (2.0.2-1) …
Setting up tree (2.0.2-1) …
Processing triggers for man-db (2.10.2-1) …
```

[2]:
```
#@title Imports

import numpy as np

import transformers
import evaluate

from datasets import load_dataset
from torchinfo import summary

from transformers import AutoTokenizer, AutoModel,␣
 ↪AutoModelForSequenceClassification
from transformers import TrainingArguments, Trainer
```

[31]:
```
# @title Mount Google Drive
```

[12]:
```
from google.colab import drive
drive.mount('/content/drive')
```

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

[13]:
```
dir_root = '/content/drive/MyDrive/266-final/'
dir_data = '/content/drive/MyDrive/266-final/data/'
dir_models = '/content/drive/MyDrive/266-final/models/'
dir_results = '/content/drive/MyDrive/266-final/results/'
```

[45]:
```
!tree -L 2 /content/drive/MyDrive/266-final/
```

```
/content/drive/MyDrive/266-final/
  data
      dev-articles
      dev-task-TC-template.out
      README.md
      train-articles
```

```
        train-labels-task1-span-identification
        train-labels-task2-technique-classification
        train-task1-SI.labels
        train-task2-TC.labels
    models
    notebook-scripts
        1 Propaganda Dataset Classification Functional Test MVP with BERT
base.ipynb
    paper
    results
    slides

10 directories, 5 files
```

[28]:
```
!ls -R /content/drive/MyDrive/266-final/
```

```
/content/drive/MyDrive/266-final/:
data  models  notebook-scripts  paper  results  slides

/content/drive/MyDrive/266-final/data:

/content/drive/MyDrive/266-final/models:

/content/drive/MyDrive/266-final/notebook-scripts:
'1 Propaganda Dataset Classification Functional Test MVP with BERT base.ipynb'

/content/drive/MyDrive/266-final/paper:

/content/drive/MyDrive/266-final/results:

/content/drive/MyDrive/266-final/slides:
```

[33]:
```
#@title Import Data
```

[43]:
```
!ls -li /content/drive/MyDrive/266-final/data/
```

```
total 375
3107 drwx------ 2 root root   4096 Apr  2 18:55 dev-articles
3112 -rw------- 1 root root  22850 Dec 11  2019 dev-task-TC-template.out
3114 -rw------- 1 root root   4886 Dec 11  2019 README.md
3108 drwx------ 2 root root   4096 Apr  2 18:55 train-articles
3109 drwx------ 2 root root   4096 Apr  2 18:55 train-labels-task1-span-
identification
3110 drwx------ 2 root root   4096 Apr  2 18:55 train-labels-task2-technique-
classification
3111 -rw------- 1 root root 108269 Dec 11  2019 train-task1-SI.labels
3113 -rw------- 1 root root 230658 Dec 11  2019 train-task2-TC.labels
```

[49]:
```
# dataset = load_dataset("sem_eval_2020_task_11") # does not work, revert to␣
↪manual load
```

[51]:
```
import os
import pandas as pd

DATA_DIR = "/content/drive/MyDrive/266-final/data/"

# Directories
TRAIN_ARTICLES_DIR = os.path.join(DATA_DIR, "train-articles")
DEV_ARTICLES_DIR   = os.path.join(DATA_DIR, "dev-articles")

# Label directories (if you need them at the directory level)
TRAIN_LABELS_TASK1_DIR = os.path.join(DATA_DIR,␣
↪"train-labels-task1-span-identification")
TRAIN_LABELS_TASK2_DIR = os.path.join(DATA_DIR,␣
↪"train-labels-task2-technique-classification")

# Individual label files
TRAIN_TASK1_LABELS_FILE = os.path.join(DATA_DIR, "train-task1-SI.labels")
TRAIN_TASK2_LABELS_FILE = os.path.join(DATA_DIR, "train-task2-TC.labels")

# Dev task template (if needed)
DEV_TC_TEMPLATE_FILE = os.path.join(DATA_DIR, "dev-task-TC-template.out")
```

[52]:
```
def load_articles_from_directory(directory):
    """
    Reads all files from the given directory and returns
    a list of dicts: [{"filename": ..., "text": ...}, ...]
    """
    articles = []
    for filename in sorted(os.listdir(directory)):
        filepath = os.path.join(directory, filename)
        if os.path.isfile(filepath) and filename.endswith(".txt"):
            with open(filepath, "r", encoding="utf-8") as f:
                text = f.read()
            articles.append({
                "filename": filename,
                "text": text
            })
    return articles

train_articles_list = load_articles_from_directory(TRAIN_ARTICLES_DIR)
dev_articles_list   = load_articles_from_directory(DEV_ARTICLES_DIR)

# Convert lists of dicts to DataFrames if desired
train_articles_df = pd.DataFrame(train_articles_list)
```

```
dev_articles_df   = pd.DataFrame(dev_articles_list)

print("Number of training articles:", len(train_articles_df))
print("Number of dev articles:", len(dev_articles_df))
```

```
Number of training articles: 371
Number of dev articles: 75
```

[53]:
```
train = train_articles_df
dev   = dev_articles_df
```

[54]:
```
# Adjust 'sep' to the correct delimiter (e.g., '\t' for TSV, ',' for CSV, etc.)
train_task1_labels = pd.read_csv(TRAIN_TASK1_LABELS_FILE, sep="\t", header=None)
train_task2_labels = pd.read_csv(TRAIN_TASK2_LABELS_FILE, sep="\t", header=None)

print(train_task1_labels.head())
print(train_task2_labels.head())
```

```
           0     1     2
0  111111111   265   323
1  111111111  1795  1935
2  111111111   149   157
3  111111111  1069  1091
4  111111111  1334  1462
           0                          1     2     3
0  111111111       Appeal_to_Authority   265   323
1  111111111       Appeal_to_Authority  1795  1935
2  111111111                    Doubt   149   157
3  111111111               Repetition  1069  1091
4  111111111  Appeal_to_fear-prejudice  1334  1462
```

[55]:
```
# Example: store the DataFrame directly
train = {
    "articles": train_articles_df,
    "task1_labels": train_task1_labels,
    "task2_labels": train_task2_labels
}

# Similarly for dev, if you have dev labels or template files:
dev_task_tc_template = pd.read_csv(DEV_TC_TEMPLATE_FILE, sep="\t", header=None)
dev = {
    "articles": dev_articles_df,
    "tc_template": dev_task_tc_template
}
```

[56]:
```
print("Training data keys:", train.keys())
print("First few train articles:\n", train["articles"].head(), "\n")
```

```
print("First few train Task1 labels:\n", train["task1_labels"].head(), "\n")
print("First few train Task2 labels:\n", train["task2_labels"].head(), "\n")

print("Dev data keys:", dev.keys())
print("First few dev articles:\n", dev["articles"].head(), "\n")
print("Dev TC template:\n", dev["tc_template"].head(), "\n")
```

```
Training data keys: dict_keys(['articles', 'task1_labels', 'task2_labels'])
First few train articles:
                filename                                               text
0  article111111111.txt  Next plague outbreak in Madagascar could be 's…
1  article111111112.txt  US bloggers banned from entering UK\n\nTwo pro…
2  article111111113.txt  Kate Steinle's death at the hands of a Mexican…
3  article111111114.txt  U.S. judge frees Indonesian immigrant held by …
4  article111111115.txt  Here are all the sexual misconduct accusations…

First few train Task1 labels:
            0     1     2
0  111111111   265   323
1  111111111  1795  1935
2  111111111   149   157
3  111111111  1069  1091
4  111111111  1334  1462

First few train Task2 labels:
            0                          1     2     3
0  111111111       Appeal_to_Authority   265   323
1  111111111       Appeal_to_Authority  1795  1935
2  111111111                    Doubt   149   157
3  111111111               Repetition  1069  1091
4  111111111  Appeal_to_fear-prejudice  1334  1462

Dev data keys: dict_keys(['articles', 'tc_template'])
First few dev articles:
                filename                                               text
0  article730081389.txt  Police had previously gone to home where Ohio …
1  article730093263.txt  America's Immigration Voice.\n\nEarlier, I blo…
2  article730246508.txt  Man arrested for allegedly buying gun used in …
3  article730269378.txt  America's Immigration Voice.\n\nThanks for pub…
4  article738028498.txt  Humanity’s WIPEOUT Foreshadowed?\nWorld Health…

Dev TC template:
            0  1     2     3
0  730093263  ?   123   128
1  730093263  ?   352   357
2  730093263  ?  1370  1393
3  730093263  ?  2434  2439
4  730093263  ?  2699  2807
```

[ ]: