FinBERTFineTuning

April 19, 2025

```
[1]: !pip install transformers
     !pip install datasets
     !pip install evaluate
     !pip cache purge
     !pip install torch==2.5.1 torchvision==0.20.1+cu124 -q
    Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-
    packages (4.51.3)
    Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
    packages (from transformers) (3.18.0)
    Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in
    /usr/local/lib/python3.11/dist-packages (from transformers) (0.30.2)
    Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-
    packages (from transformers) (2.0.2)
    Requirement already satisfied: packaging>=20.0 in
    /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
    Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-
    packages (from transformers) (6.0.2)
    Requirement already satisfied: regex!=2019.12.17 in
    /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.6)
    Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-
    packages (from transformers) (2.32.3)
    Requirement already satisfied: tokenizers<0.22,>=0.21 in
    /usr/local/lib/python3.11/dist-packages (from transformers) (0.21.1)
    Requirement already satisfied: safetensors>=0.4.3 in
    /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
    Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-
    packages (from transformers) (4.67.1)
    Requirement already satisfied: fsspec>=2023.5.0 in
    /usr/local/lib/python3.11/dist-packages (from huggingface-
    hub<1.0,>=0.30.0->transformers) (2025.3.2)
    Requirement already satisfied: typing-extensions>=3.7.4.3 in
    /usr/local/lib/python3.11/dist-packages (from huggingface-
    hub<1.0,>=0.30.0->transformers) (4.13.2)
    Requirement already satisfied: charset-normalizer<4,>=2 in
    /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4.1)
    Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
```

```
packages (from requests->transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests->transformers)
(2025.1.31)
Collecting datasets
 Downloading datasets-3.5.0-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from datasets) (3.18.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-
packages (from datasets) (2.0.2)
Requirement already satisfied: pyarrow>=15.0.0 in
/usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
 Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages
(from datasets) (2.2.2)
Requirement already satisfied: requests>=2.32.2 in
/usr/local/lib/python3.11/dist-packages (from datasets) (2.32.3)
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.11/dist-
packages (from datasets) (4.67.1)
Collecting xxhash (from datasets)
 Downloading
xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
(12 kB)
Collecting multiprocess<0.70.17 (from datasets)
  Downloading multiprocess-0.70.16-py311-none-any.whl.metadata (7.2 kB)
Collecting fsspec<=2024.12.0,>=2023.1.0 (from
fsspec[http]<=2024.12.0,>=2023.1.0->datasets)
  Downloading fsspec-2024.12.0-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-
packages (from datasets) (3.11.15)
Requirement already satisfied: huggingface-hub>=0.24.0 in
/usr/local/lib/python3.11/dist-packages (from datasets) (0.30.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-
packages (from datasets) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-
packages (from datasets) (6.0.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-
packages (from aiohttp->datasets) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
```

```
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (6.4.3)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (0.3.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.19.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.24.0->datasets)
(4.13.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests>=2.32.2->datasets) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(2025.1.31)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas->datasets) (1.17.0)
Downloading datasets-3.5.0-py3-none-any.whl (491 kB)
                         491.2/491.2 kB
21.4 MB/s eta 0:00:00
Downloading dill-0.3.8-py3-none-any.whl (116 kB)
                         116.3/116.3 kB
12.5 MB/s eta 0:00:00
Downloading fsspec-2024.12.0-py3-none-any.whl (183 kB)
                         183.9/183.9 kB
14.6 MB/s eta 0:00:00
Downloading multiprocess-0.70.16-py311-none-any.whl (143 kB)
                         143.5/143.5 kB
14.2 MB/s eta 0:00:00
Downloading
xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
                         194.8/194.8 kB
12.5 MB/s eta 0:00:00
Installing collected packages: xxhash, fsspec, dill, multiprocess,
datasets
  Attempting uninstall: fsspec
   Found existing installation: fsspec 2025.3.2
   Uninstalling fsspec-2025.3.2:
```

Successfully uninstalled fsspec-2025.3.2

```
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.
```

gcsfs 2025.3.2 requires fsspec==2025.3.2, but you have fsspec 2024.12.0 which is incompatible.

- 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.

Successfully installed datasets-3.5.0 dill-0.3.8 fsspec-2024.12.0

```
multiprocess-0.70.16 xxhash-3.5.0
Collecting evaluate
  Downloading evaluate-0.4.3-py3-none-any.whl.metadata (9.2 kB)
Requirement already satisfied: datasets>=2.0.0 in
/usr/local/lib/python3.11/dist-packages (from evaluate) (3.5.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-
packages (from evaluate) (2.0.2)
Requirement already satisfied: dill in /usr/local/lib/python3.11/dist-packages
(from evaluate) (0.3.8)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages
(from evaluate) (2.2.2)
Requirement already satisfied: requests>=2.19.0 in
/usr/local/lib/python3.11/dist-packages (from evaluate) (2.32.3)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.11/dist-
packages (from evaluate) (4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages
(from evaluate) (3.5.0)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.11/dist-
packages (from evaluate) (0.70.16)
Requirement already satisfied: fsspec>=2021.05.0 in
/usr/local/lib/python3.11/dist-packages (from fsspec[http]>=2021.05.0->evaluate)
(2024.12.0)
Requirement already satisfied: huggingface-hub>=0.7.0 in
/usr/local/lib/python3.11/dist-packages (from evaluate) (0.30.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-
packages (from evaluate) (24.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from datasets>=2.0.0->evaluate) (3.18.0)
Requirement already satisfied: pyarrow>=15.0.0 in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.0.0->evaluate)
(18.1.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-
packages (from datasets>=2.0.0->evaluate) (3.11.15)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-
packages (from datasets>=2.0.0->evaluate) (6.0.2)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.7.0->evaluate)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests>=2.19.0->evaluate) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate)
```

```
(2025.1.31)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas->evaluate) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-
packages (from pandas->evaluate) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas->evaluate) (2025.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-
packages (from aiohttp->datasets>=2.0.0->evaluate) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (6.4.3)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (0.3.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (1.19.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas->evaluate) (1.17.0)
Downloading evaluate-0.4.3-py3-none-any.whl (84 kB)
                         84.0/84.0 kB
8.1 MB/s eta 0:00:00
Installing collected packages: evaluate
Successfully installed evaluate-0.4.3
Files removed: 36
ERROR: Ignored the following yanked versions: 0.1.6, 0.1.7, 0.1.8, 0.1.9,
0.2.0, 0.2.1, 0.2.2, 0.2.2.post2, 0.2.2.post3, 0.15.0
ERROR: Could not find a version that satisfies the requirement
torchvision==0.20.1+cu124 (from versions: 0.15.1, 0.15.2, 0.16.0, 0.16.1,
0.16.2, 0.17.0, 0.17.1, 0.17.2, 0.18.0, 0.18.1, 0.19.0, 0.19.1, 0.20.0, 0.20.1,
0.21.0)
ERROR: No matching distribution found for
torchvision==0.20.1+cu124
```

• Here taking module which are compatible to finbert model.

1 Libraries

```
[2]: import pandas as pd
from datasets import Dataset
import transformers
from transformers import AutoTokenizer, AutoModelForSequenceClassification,

→TrainingArguments, Trainer, pipeline
from sklearn.metrics import classification_report, confusion_matrix,

→ConfusionMatrixDisplay
import matplotlib.pyplot as plt
```

2 Dataset Loading, Preprocessing, and Splitting

```
[3]: df = pd.read_csv('all-data.csv', encoding='ISO-8859-1', header=None)
    df.columns = ['sentiment', 'text']
    df.dropna(inplace=True)

label2id = {'negative': 0, 'neutral': 1, 'positive': 2}
    df['label'] = df['sentiment'].map(label2id)

# Hugging Face Dataset
    hf_dataset = Dataset.from_pandas(df[['text', 'label']])
    hf_dataset = hf_dataset.train_test_split(test_size=0.2)
```

3 Tokenization of Dataset Using FinBERT Tokenizer

```
[4]: tokenizer = AutoTokenizer.from_pretrained("yiyanghkust/finbert-tone")

# Tokenize the dataset
def tokenize_function(examples):
    return tokenizer(examples["text"], padding="max_length", truncation=True)

tokenized_datasets = hf_dataset.map(tokenize_function, batched=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), 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

4 Loading the FinBERT Model for Sequence Classification

| 0/970 [00:00<?, ? examples/s]

Map:

0%1

```
[5]: model = AutoModelForSequenceClassification.from_pretrained(
    "yiyanghkust/finbert-tone",
    num_labels=3,
    id2label={0: "negative", 1: "neutral", 2: "positive"},
    label2id={"negative": 0, "neutral": 1, "positive": 2}
)

pytorch_model.bin: 0%|    | 0.00/439M [00:00<?, ?B/s]</pre>
```

5 Setting Up Training Arguments for Fine-Tuning

```
[6]: training_args = TrainingArguments(
    output_dir="./finbert-finetuned",
    num_train_epochs=4,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    learning_rate=2e-5,
    weight_decay=0.01,
    logging_dir="./logs",
    report_to="none"
)
```

6 Initializing the Trainer for Model Fine-Tuning

```
tokenizer=tokenizer
     )
    <ipython-input-7-eca663cc33b8>:1: FutureWarning: `tokenizer` is deprecated and
    will be removed in version 5.0.0 for `Trainer.__init__`. Use `processing_class`
    instead.
      trainer = Trainer(
[8]: trainer.train()
                         0%1
    model.safetensors:
                                      | 0.00/439M [00:00<?, ?B/s]
    <IPython.core.display.HTML object>
[8]: TrainOutput(global_step=972, training_loss=0.338934066364304,
    metrics={'train_runtime': 196.9126, 'train_samples_per_second': 78.735,
     'train_steps_per_second': 4.936, 'total_flos': 461973716617392.0, 'train_loss':
     0.338934066364304, 'epoch': 4.0})
        Evaluating the Model and Generating Predictions
[9]: results = trainer.evaluate()
     print("Evaluation results:")
     for key, value in results.items():
        print(f"{key}: {value}")
     # calculating predictions on test set
     predictions = trainer.predict(tokenized_datasets["test"])
     preds = predictions.predictions.argmax(-1)
     print("Classification report on Test data")
     print(classification_report(tokenized_datasets["test"]["label"], preds,
                                 target_names=["negative", "neutral", "positive"]))
    <IPython.core.display.HTML object>
    Evaluation results:
    eval_loss: 0.6462458968162537
    eval_runtime: 3.3138
    eval_samples_per_second: 292.715
    eval_steps_per_second: 18.408
    epoch: 4.0
    Classification report on Test data
                  precision
                               recall f1-score
                                                  support
                       0.75
                                 0.84
                                           0.80
                                                      120
        negative
                       0.90
                                 0.86
                                           0.88
                                                      586
         neutral
        positive
                       0.78
                                 0.83
                                           0.80
                                                      264
```

```
accuracy 0.85 970 macro avg 0.81 0.84 0.83 970 weighted avg 0.85 0.85 0.85 970
```

8 Saving the Fine-Tuned Model and Tokenizer

9 Manual testing for News headline

10 Confusion Matrix for Model Evaluation

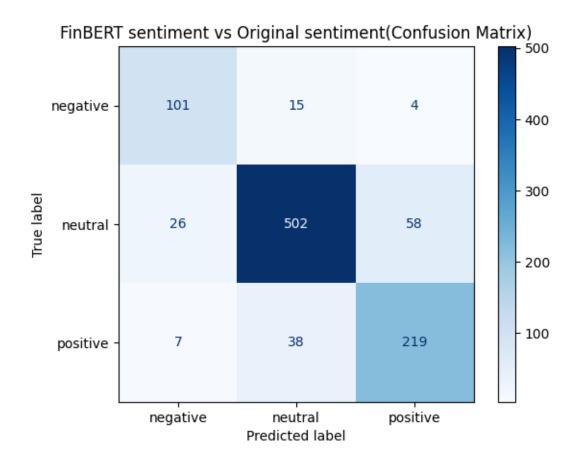
```
[13]: finbert_custom = pipeline(
    "text-classification",
    model="./my-finbert-finetuned",
    tokenizer=tokenizer,
    device=0
)

test_df = tokenized_datasets["test"].to_pandas()
def get_custom_sentiment(text):
    try:
        result = finbert_custom(text[:512])
```

Device set to use cuda:0

You seem to be using the pipelines sequentially on GPU. In order to maximize efficiency please use a dataset

<Figure size 600x500 with 0 Axes>



11 To download Fine-tuned model on local pc

```
adding: my-finbert-finetuned/ (stored 0%)
adding: my-finbert-finetuned/ tokenizer_config.json (deflated 74%)
adding: my-finbert-finetuned/model.safetensors (deflated 7%)
adding: my-finbert-finetuned/model.safetensors (deflated 42%)
adding: my-finbert-finetuned/special_tokens_map.json (deflated 42%)
adding: my-finbert-finetuned/config.json (deflated 52%)
adding: my-finbert-finetuned/training_args.bin (deflated 52%)
adding: my-finbert-finetuned/vocab.txt (deflated 50%)
adding: my-finbert-finetuned/tokenizer.json (deflated 70%)

[15]: from google.colab import files
files.download("my-finbert-finetuned.zip")

<IPython.core.display.Javascript object>
<IPython.core.display.Javascript object>
```