Transformer_Model_Training&Testing&Evaluation

April 21, 2025

0.0.1 Install Required Libraries

This cell installs all the necessary Python libraries for the notebook: - `datasets`: For loading and processing datasets. - `transformers`: For using pre-trained models and tokenizers. - `seqeval`: For evaluating NER models. - `gdown`: For downloading files from Google Drive. - `pandas`, `scikit-learn`: For data manipulation and evaluation. - `torch`: For PyTorch-based model training. - `openai`: For interacting with OpenAI's API (if needed).

[ ]:
```
!pip install datasets
!pip install -U accelerate
!pip install -U transformers
!pip install seqeval
!pip install gdown
!pip install pandas scikit-learn
!pip install torch
!pip install openai
```

```
Collecting datasets
  Downloading datasets-3.4.1-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from datasets) (3.17.0)
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)
```

```
Requirement already satisfied: fsspec<=2024.12.0,>=2023.1.0 in
/usr/local/lib/python3.11/dist-packages (from
fsspec[http]<=2024.12.0,>=2023.1.0->datasets) (2024.10.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-
packages (from datasets) (3.11.13)
Requirement already satisfied: huggingface-hub>=0.24.0 in
/usr/local/lib/python3.11/dist-packages (from datasets) (0.28.1)
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) (6.1.0)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (0.3.0)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.18.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.24.0->datasets)
(4.12.2)
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) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.1)
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-3.4.1-py3-none-any.whl (487 kB)
```

```
                         487.4/487.4 kB
14.1 MB/s eta 0:00:00
Downloading dill-0.3.8-py3-none-any.whl (116 kB)
                         116.3/116.3 kB
5.2 MB/s eta 0:00:00
Downloading multiprocess-0.70.16-py311-none-any.whl (143 kB)
                         143.5/143.5 kB
5.4 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
7.6 MB/s eta 0:00:00
Installing collected packages: xxhash, dill, multiprocess, datasets
Successfully installed datasets-3.4.1 dill-0.3.8 multiprocess-0.70.16
xxhash-3.5.0
Requirement already satisfied: accelerate in /usr/local/lib/python3.11/dist-
packages (1.3.0)
Collecting accelerate
  Downloading accelerate-1.5.2-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: numpy<3.0.0,>=1.17 in
/usr/local/lib/python3.11/dist-packages (from accelerate) (2.0.2)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.11/dist-packages (from accelerate) (24.2)
Requirement already satisfied: psutil in /usr/local/lib/python3.11/dist-packages
(from accelerate) (5.9.5)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.11/dist-packages
(from accelerate) (6.0.2)
Requirement already satisfied: torch>=2.0.0 in /usr/local/lib/python3.11/dist-
packages (from accelerate) (2.6.0+cu124)
Requirement already satisfied: huggingface-hub>=0.21.0 in
/usr/local/lib/python3.11/dist-packages (from accelerate) (0.28.1)
Requirement already satisfied: safetensors>=0.4.3 in
/usr/local/lib/python3.11/dist-packages (from accelerate) (0.5.3)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from huggingface-hub>=0.21.0->accelerate) (3.17.0)
Requirement already satisfied: fsspec>=2023.5.0 in
/usr/local/lib/python3.11/dist-packages (from huggingface-
hub>=0.21.0->accelerate) (2024.10.0)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-
packages (from huggingface-hub>=0.21.0->accelerate) (2.32.3)
Requirement already satisfied: tqdm>=4.42.1 in /usr/local/lib/python3.11/dist-
packages (from huggingface-hub>=0.21.0->accelerate) (4.67.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-
hub>=0.21.0->accelerate) (4.12.2)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-
packages (from torch>=2.0.0->accelerate) (3.4.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages
```

```
(from torch>=2.0.0->accelerate) (3.1.6)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch>=2.0.0->accelerate)
  Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch>=2.0.0->accelerate)
  Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch>=2.0.0->accelerate)
  Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch>=2.0.0->accelerate)
  Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch>=2.0.0->accelerate)
  Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufft-cu12==11.2.1.3 (from torch>=2.0.0->accelerate)
  Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.5.147 (from torch>=2.0.0->accelerate)
  Downloading nvidia_curand_cu12-10.3.5.147-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch>=2.0.0->accelerate)
  Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparse-cu12==12.3.1.170 (from torch>=2.0.0->accelerate)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate) (0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.0.0->accelerate)
(12.4.127)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch>=2.0.0->accelerate)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-
packages (from torch>=2.0.0->accelerate) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-
packages (from torch>=2.0.0->accelerate) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.11/dist-packages (from
sympy==1.13.1->torch>=2.0.0->accelerate) (1.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2->torch>=2.0.0->accelerate)
(3.0.2)
```

```
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests->huggingface-
hub>=0.21.0->accelerate) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests->huggingface-hub>=0.21.0->accelerate) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests->huggingface-
hub>=0.21.0->accelerate) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests->huggingface-
hub>=0.21.0->accelerate) (2025.1.31)
Downloading accelerate-1.5.2-py3-none-any.whl (345 kB)
                         345.1/345.1 kB
16.0 MB/s eta 0:00:00
Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl
(363.4 MB)
                         363.4/363.4 MB
5.9 MB/s eta 0:00:00
Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl (13.8 MB)
                         13.8/13.8 MB
105.8 MB/s eta 0:00:00
Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl (24.6 MB)
                         24.6/24.6 MB
84.2 MB/s eta 0:00:00
Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl (883 kB)
                         883.7/883.7 kB
60.1 MB/s eta 0:00:00
Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl
(664.8 MB)
                         664.8/664.8 MB
2.8 MB/s eta 0:00:00
Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl
(211.5 MB)
                         211.5/211.5 MB
5.7 MB/s eta 0:00:00
Downloading nvidia_curand_cu12-10.3.5.147-py3-none-
manylinux2014_x86_64.whl (56.3 MB)
                         56.3/56.3 MB
15.0 MB/s eta 0:00:00
Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-
manylinux2014_x86_64.whl (127.9 MB)
                         127.9/127.9 MB
7.6 MB/s eta 0:00:00
Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-
manylinux2014_x86_64.whl (207.5 MB)
```

```
                         207.5/207.5 MB
6.3 MB/s eta 0:00:00
Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl (21.1 MB)
                         21.1/21.1 MB
87.5 MB/s eta 0:00:00
Installing collected packages: nvidia-nvjitlink-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, accelerate
  Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.5.82
    Uninstalling nvidia-nvjitlink-cu12-12.5.82:
      Successfully uninstalled nvidia-nvjitlink-cu12-12.5.82
  Attempting uninstall: nvidia-curand-cu12
    Found existing installation: nvidia-curand-cu12 10.3.6.82
    Uninstalling nvidia-curand-cu12-10.3.6.82:
      Successfully uninstalled nvidia-curand-cu12-10.3.6.82
  Attempting uninstall: nvidia-cufft-cu12
    Found existing installation: nvidia-cufft-cu12 11.2.3.61
    Uninstalling nvidia-cufft-cu12-11.2.3.61:
      Successfully uninstalled nvidia-cufft-cu12-11.2.3.61
  Attempting uninstall: nvidia-cuda-runtime-cu12
    Found existing installation: nvidia-cuda-runtime-cu12 12.5.82
    Uninstalling nvidia-cuda-runtime-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-runtime-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-nvrtc-cu12
    Found existing installation: nvidia-cuda-nvrtc-cu12 12.5.82
    Uninstalling nvidia-cuda-nvrtc-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-cupti-cu12
    Found existing installation: nvidia-cuda-cupti-cu12 12.5.82
    Uninstalling nvidia-cuda-cupti-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-cupti-cu12-12.5.82
  Attempting uninstall: nvidia-cublas-cu12
    Found existing installation: nvidia-cublas-cu12 12.5.3.2
    Uninstalling nvidia-cublas-cu12-12.5.3.2:
      Successfully uninstalled nvidia-cublas-cu12-12.5.3.2
  Attempting uninstall: nvidia-cusparse-cu12
    Found existing installation: nvidia-cusparse-cu12 12.5.1.3
    Uninstalling nvidia-cusparse-cu12-12.5.1.3:
      Successfully uninstalled nvidia-cusparse-cu12-12.5.1.3
  Attempting uninstall: nvidia-cudnn-cu12
    Found existing installation: nvidia-cudnn-cu12 9.3.0.75
    Uninstalling nvidia-cudnn-cu12-9.3.0.75:
      Successfully uninstalled nvidia-cudnn-cu12-9.3.0.75
  Attempting uninstall: nvidia-cusolver-cu12
    Found existing installation: nvidia-cusolver-cu12 11.6.3.83
```

```
    Uninstalling nvidia-cusolver-cu12-11.6.3.83:
      Successfully uninstalled nvidia-cusolver-cu12-11.6.3.83
  Attempting uninstall: accelerate
    Found existing installation: accelerate 1.3.0
    Uninstalling accelerate-1.3.0:
      Successfully uninstalled accelerate-1.3.0
Successfully installed accelerate-1.5.2 nvidia-cublas-cu12-12.4.5.8 nvidia-cuda-
cupti-cu12-12.4.127 nvidia-cuda-nvrtc-cu12-12.4.127 nvidia-cuda-runtime-
cu12-12.4.127 nvidia-cudnn-cu12-9.1.0.70 nvidia-cufft-cu12-11.2.1.3 nvidia-
curand-cu12-10.3.5.147 nvidia-cusolver-cu12-11.6.1.9 nvidia-cusparse-
cu12-12.3.1.170 nvidia-nvjitlink-cu12-12.4.127
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-
packages (4.48.3)
Collecting transformers
  Downloading transformers-4.49.0-py3-none-any.whl.metadata (44 kB)
                           44.0/44.0 kB
4.1 MB/s eta 0:00:00
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from transformers) (3.17.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.26.0 in
/usr/local/lib/python3.11/dist-packages (from transformers) (0.28.1)
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>=0.4.1 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,>=0.26.0->transformers) (2024.10.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-
hub<1.0,>=0.26.0->transformers) (4.12.2)
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)
Downloading transformers-4.49.0-py3-none-any.whl (10.0 MB)
                        10.0/10.0 MB
108.9 MB/s eta 0:00:00
Installing collected packages: transformers
  Attempting uninstall: transformers
    Found existing installation: transformers 4.48.3
    Uninstalling transformers-4.48.3:
      Successfully uninstalled transformers-4.48.3
Successfully installed transformers-4.49.0
Collecting seqeval
  Downloading seqeval-1.2.2.tar.gz (43 kB)
                           43.6/43.6 kB
3.8 MB/s eta 0:00:00
  Preparing metadata (setup.py) … done
Requirement already satisfied: numpy>=1.14.0 in /usr/local/lib/python3.11/dist-
packages (from seqeval) (2.0.2)
Requirement already satisfied: scikit-learn>=0.21.3 in
/usr/local/lib/python3.11/dist-packages (from seqeval) (1.6.1)
Requirement already satisfied: scipy>=1.6.0 in /usr/local/lib/python3.11/dist-
packages (from scikit-learn>=0.21.3->seqeval) (1.14.1)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-
packages (from scikit-learn>=0.21.3->seqeval) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in
/usr/local/lib/python3.11/dist-packages (from scikit-learn>=0.21.3->seqeval)
(3.6.0)
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=16161
sha256=351ec571a05a0fe51bfcad0b2b3ae90d79ebaf2525958ef85ce1796f504d8399
  Stored in directory: /root/.cache/pip/wheels/bc/92/f0/243288f899c2eacdfa8c5f9a
ede4c71a9bad0ee26a01dc5ead
Successfully built seqeval
Installing collected packages: seqeval
Successfully installed seqeval-1.2.2
Requirement already satisfied: gdown in /usr/local/lib/python3.11/dist-packages
(5.2.0)
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.11/dist-
packages (from gdown) (4.13.3)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from gdown) (3.17.0)
Requirement already satisfied: requests[socks] in
/usr/local/lib/python3.11/dist-packages (from gdown) (2.32.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages
(from gdown) (4.67.1)
```

```
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.11/dist-
packages (from beautifulsoup4->gdown) (2.6)
Requirement already satisfied: typing-extensions>=4.0.0 in
/usr/local/lib/python3.11/dist-packages (from beautifulsoup4->gdown) (4.12.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests[socks]->gdown) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests[socks]->gdown) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests[socks]->gdown) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests[socks]->gdown)
(2025.1.31)
Requirement already satisfied: PySocks!=1.5.7,>=1.5.6 in
/usr/local/lib/python3.11/dist-packages (from requests[socks]->gdown) (1.7.1)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages
(2.2.2)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-
packages (1.6.1)
Requirement already satisfied: numpy>=1.23.2 in /usr/local/lib/python3.11/dist-
packages (from pandas) (2.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-
packages (from pandas) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas) (2025.1)
Requirement already satisfied: scipy>=1.6.0 in /usr/local/lib/python3.11/dist-
packages (from scikit-learn) (1.14.1)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-
packages (from scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in
/usr/local/lib/python3.11/dist-packages (from scikit-learn) (3.6.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
Requirement already satisfied: torch in /usr/local/lib/python3.11/dist-packages
(2.6.0+cu124)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from torch) (3.17.0)
Requirement already satisfied: typing-extensions>=4.10.0 in
/usr/local/lib/python3.11/dist-packages (from torch) (4.12.2)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-
packages (from torch) (3.4.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages
(from torch) (3.1.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages
(from torch) (2024.10.0)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in
```

```
/usr/local/lib/python3.11/dist-packages (from torch) (12.4.127)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch) (12.4.127)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch) (12.4.127)
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in
/usr/local/lib/python3.11/dist-packages (from torch) (9.1.0.70)
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in
/usr/local/lib/python3.11/dist-packages (from torch) (12.4.5.8)
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in
/usr/local/lib/python3.11/dist-packages (from torch) (11.2.1.3)
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in
/usr/local/lib/python3.11/dist-packages (from torch) (10.3.5.147)
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in
/usr/local/lib/python3.11/dist-packages (from torch) (11.6.1.9)
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in
/usr/local/lib/python3.11/dist-packages (from torch) (12.3.1.170)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in
/usr/local/lib/python3.11/dist-packages (from torch) (0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in
/usr/local/lib/python3.11/dist-packages (from torch) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch) (12.4.127)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in
/usr/local/lib/python3.11/dist-packages (from torch) (12.4.127)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-
packages (from torch) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-
packages (from torch) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch) (1.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2->torch) (3.0.2)
Requirement already satisfied: openai in /usr/local/lib/python3.11/dist-packages
(1.61.1)
Requirement already satisfied: anyio<5,>=3.5.0 in
/usr/local/lib/python3.11/dist-packages (from openai) (3.7.1)
Requirement already satisfied: distro<2,>=1.7.0 in
/usr/local/lib/python3.11/dist-packages (from openai) (1.9.0)
Requirement already satisfied: httpx<1,>=0.23.0 in
/usr/local/lib/python3.11/dist-packages (from openai) (0.28.1)
Requirement already satisfied: jiter<1,>=0.4.0 in
/usr/local/lib/python3.11/dist-packages (from openai) (0.9.0)
Requirement already satisfied: pydantic<3,>=1.9.0 in
/usr/local/lib/python3.11/dist-packages (from openai) (2.10.6)
Requirement already satisfied: sniffio in /usr/local/lib/python3.11/dist-
packages (from openai) (1.3.1)
Requirement already satisfied: tqdm>4 in /usr/local/lib/python3.11/dist-packages
```

```
(from openai) (4.67.1)
Requirement already satisfied: typing-extensions<5,>=4.11 in
/usr/local/lib/python3.11/dist-packages (from openai) (4.12.2)
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.11/dist-
packages (from anyio<5,>=3.5.0->openai) (3.10)
Requirement already satisfied: certifi in /usr/local/lib/python3.11/dist-
packages (from httpx<1,>=0.23.0->openai) (2025.1.31)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-
packages (from httpx<1,>=0.23.0->openai) (1.0.7)
Requirement already satisfied: h11<0.15,>=0.13 in
/usr/local/lib/python3.11/dist-packages (from
httpcore==1.*->httpx<1,>=0.23.0->openai) (0.14.0)
Requirement already satisfied: annotated-types>=0.6.0 in
/usr/local/lib/python3.11/dist-packages (from pydantic<3,>=1.9.0->openai)
(0.7.0)
Requirement already satisfied: pydantic-core==2.27.2 in
/usr/local/lib/python3.11/dist-packages (from pydantic<3,>=1.9.0->openai)
(2.27.2)
```

0.0.2 Import Libraries

This cell imports all the required libraries and modules: - `transformers`: For model training, tokenization, and evaluation. - `datasets`: For loading and processing datasets. - `ast`: For safely evaluating strings as Python objects. - `gdown`: For downloading files from Google Drive. - `pandas`: For data manipulation. - `torch`: For PyTorch-based model training. - `openai`: For interacting with OpenAI's API (if needed). - `google.colab.drive`: For mounting Google Drive.

[ ]:
```
from transformers import (
    AutoModelForTokenClassification, AutoTokenizer,␣
 ↪DataCollatorForTokenClassification,
    TrainingArguments, Trainer
)
from datasets import load_dataset
import ast
import gdown
import pandas as pd
import torch
import openai
```

0.0.3 Mount Google Drive

Mount your google drive to save the datasets, model over the drive.

Note: If you want to run the code locally, update the file paths accordingly for loading and saving datasets and models.

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

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

0.0.4 Download Dataset Files

This cell downloads the training, validation, and test datasets from Google Drive using `gdown`. The datasets are saved as CSV files: - `train.csv`: Training dataset. - `val.csv`: Validation dataset. - `test.csv`: Test dataset.

0.0.5 Load Tokenizer

This cell loads the tokenizer for the `bert-base-cased` model from Hugging Face's `transformers` library. The tokenizer is used to preprocess text data for the model.

0.0.6 Load Dataset

This cell loads the training and validation datasets from the CSV files using the `datasets` library. The dataset is stored in a `DatasetDict` object.

[ ]:
```
# https://drive.google.com/file/d/14RDeg4gRMhAzxgb3oB8uJ5-JT2w24_Tp/view?
 ↪usp=sharing
train_file_id = "14RDeg4gRMhAzxgb3oB8uJ5-JT2w24_Tp"
# https://drive.google.com/file/d/15BOK8cly_iY3ywGPrwaqmGBhAGDlwYqR/view?
 ↪usp=sharing
val_file_id = "15BOK8cly_iY3ywGPrwaqmGBhAGDlwYqR"
# https://drive.google.com/file/d/1EUmyd3w0lVIG-4rECFMqPIL4tjGemtIL/view?
 ↪usp=sharing
test_file_id = "1EUmyd3w0lVIG-4rECFMqPIL4tjGemtIL"

gdown.download(f"https://drive.google.com/uc?id={train_file_id}", "train.csv",␣
 ↪quiet=False)
gdown.download(f"https://drive.google.com/uc?id={val_file_id}", "val.csv",␣
 ↪quiet=False)
gdown.download(f"https://drive.google.com/uc?id={test_file_id}", "test.csv",␣
 ↪quiet=False)

MODEL_NAME = "bert-base-cased"
tokenizer = AutoTokenizer.from_pretrained(MODEL_NAME, use_fast=True)

DATA_FILES = {"train": "train.csv", "val": "val.csv"}
dataset = load_dataset("csv", data_files=DATA_FILES)
```

```
Downloading…
From: https://drive.google.com/uc?id=14RDeg4gRMhAzxgb3oB8uJ5-JT2w24_Tp
To: /content/train.csv
100%|      | 14.6M/14.6M [00:00<00:00, 41.9MB/s]
Downloading…
From: https://drive.google.com/uc?id=15BOK8cly_iY3ywGPrwaqmGBhAGDlwYqR
To: /content/val.csv
100%|      | 1.82M/1.82M [00:00<00:00, 116MB/s]
```

```
Downloading…
From: https://drive.google.com/uc?id=1EUmyd3w0lVIG-4rECFMqPIL4tjGemtIL
To: /content/test.csv
100%|      | 1.83M/1.83M [00:00<00:00, 188MB/s]
/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
public models or datasets.
  warnings.warn(
```

```
tokenizer_config.json:   0%|          | 0.00/49.0 [00:00<?, ?B/s]
```

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

```
vocab.txt:   0%|          | 0.00/213k [00:00<?, ?B/s]
```

```
tokenizer.json:   0%|          | 0.00/436k [00:00<?, ?B/s]
```

```
Generating train split: 0 examples [00:00, ? examples/s]
```

```
Generating val split: 0 examples [00:00, ? examples/s]
```

0.0.7 Define Label Mappings

This cell defines the list of NER labels (`LABEL_LIST`) and creates mappings between labels and their corresponding IDs (`label2id` and `id2label`).

[ ]:
```python
LABEL_LIST = ["O", "B-PER", "I-PER", "B-EMAIL", "I-EMAIL"]
label2id = {label: i for i, label in enumerate(LABEL_LIST)}
id2label = {i: label for label, i in label2id.items()}
```

0.0.8 Convert String Columns to Lists

This cell converts the `tokens` and `ner_tags` columns from strings to Python lists using `ast.literal_eval`. This is necessary because the CSV files store these columns as strings.

[ ]:
```python
def convert_str_to_list(example):
    example["tokens"] = ast.literal_eval(example["tokens"])
    example["ner_tags"] = ast.literal_eval(example["ner_tags"])
    return example

dataset = dataset.map(convert_str_to_list)
```

```
Map:   0%|          | 0/22812 [00:00<?, ? examples/s]
```

```
Map:   0%|          | 0/2852 [00:00<?, ? examples/s]
```

0.0.9 Tokenize and Align Labels

This cell tokenizes the text data and aligns the NER labels with the tokenized input. It ensures that the labels are correctly assigned to each token, even after tokenization.

0.0.10 Tokenize Datasets

This cell applies the `tokenize_and_align_labels` function to the training and validation datasets. The tokenized datasets are stored in `tokenized_datasets`.

[ ]:
```python
def tokenize_and_align_labels(batch):
    tokenized_inputs = tokenizer(
        batch["tokens"],
        is_split_into_words=True,
        truncation=True,
        padding="max_length",
        max_length=128
    )

    all_labels = []
    for i, ner_tags in enumerate(batch["ner_tags"]):
        labels = [label2id.get(tag, label2id["O"]) for tag in ner_tags]
        word_ids = tokenized_inputs.word_ids(batch_index=i)
        if word_ids is None:
            raise ValueError("word_ids is None.")

        previous_word_idx = None
        aligned_labels = []
        for word_idx in word_ids:
            if word_idx is None:
                aligned_labels.append(-100)
            elif word_idx != previous_word_idx:
                aligned_labels.append(labels[word_idx])
            else:
                aligned_labels.append(-100)
            previous_word_idx = word_idx

        aligned_labels = aligned_labels[:128]
        aligned_labels += [-100] * (128 - len(aligned_labels))

        all_labels.append(aligned_labels)

    tokenized_inputs["labels"] = all_labels
    return tokenized_inputs

tokenized_datasets = dataset.map(tokenize_and_align_labels, batched=True)
```

```
Map:   0%|          | 0/22812 [00:00<?, ? examples/s]
```

```
Map:   0%|          | 0/2852 [00:00<?, ? examples/s]
```

0.0.11 Define Evaluation Metrics

This cell defines the `compute_metrics` function, which calculates evaluation metrics such as accuracy, precision, recall, F1-score, FPR, and FNR for the NER model.

```
[ ]: from seqeval.metrics import accuracy_score, f1_score, precision_score,␣
      ↪recall_score, classification_report
     import numpy as np

     def compute_metrics(p):
         predictions, labels = p
         predictions = np.argmax(predictions, axis=2)

         true_predictions = [
             [id2label[p] for p, l in zip(prediction, label) if l != -100]
             for prediction, label in zip(predictions, labels)
         ]
         true_labels = [
             [id2label[l] for p, l in zip(prediction, label) if l != -100]
             for prediction, label in zip(predictions, labels)
         ]

         accuracy = accuracy_score(true_labels, true_predictions)
         f1 = f1_score(true_labels, true_predictions, zero_division=0)
         precision = precision_score(true_labels, true_predictions, zero_division=0)
         recall = recall_score(true_labels, true_predictions, zero_division=0)

         report = classification_report(true_labels, true_predictions,␣
      ↪output_dict=True, zero_division=0)

         fpr = {}
         fnr = {}
         for entity_type, metrics in report.items():
             if entity_type not in ["micro avg", "macro avg", "weighted avg"]:
                 tp = metrics["support"] * metrics["recall"]
                 fn = metrics["support"] * (1 - metrics["recall"])
                 fp = metrics["support"] * (1 - metrics["precision"])

                 denominator_fpr = fp + tp + fn
                 denominator_fnr = fn + tp

                 if denominator_fpr > 0:
                     fpr[entity_type] = fp / denominator_fpr
                 else:
                     fpr[entity_type] = 0
```

```
            if denominator_fnr > 0:
                fnr[entity_type] = fn / denominator_fnr
            else:
                fnr[entity_type] = 0

    valid_fpr = [v for v in fpr.values() if not np.isnan(v)]
    valid_fnr = [v for v in fnr.values() if not np.isnan(v)]

    micro_fpr = sum(valid_fpr) / len(valid_fpr) if valid_fpr else 0
    micro_fnr = sum(valid_fnr) / len(valid_fnr) if valid_fnr else 0

    return {
        "accuracy": accuracy,
        "f1": f1,
        "precision": precision,
        "recall": recall,
        "fpr": micro_fpr,
        "fnr": micro_fnr,
        "classification_report": report,
    }
```

0.0.12 Set Environment Variables

This cell sets environment variables to configure the Hugging Face Hub download timeout and disable Weights & Biases (W&B) logging.

0.0.13 Load Pre-trained Model

This cell loads the pre-trained `bert-base-cased` model for token classification. The model is configured with the number of labels and label mappings.

0.0.14 Define Training Arguments

This cell defines the training arguments for the `Trainer` class, including: - Output directory. - Evaluation and save strategies. - Learning rate. - Batch size. - Number of epochs. - Weight decay.

0.0.15 Initialize Trainer

This cell initializes the `Trainer` class with the model, training arguments, datasets, tokenizer, data collator, and evaluation metrics.

[ ]:
```
import os
os.environ["HF_HUB_DOWNLOAD_TIMEOUT"] = "600"
os.environ["WANDB_DISABLED"] = "true"
model = AutoModelForTokenClassification.from_pretrained(
    MODEL_NAME,
    num_labels=len(LABEL_LIST),
    id2label=id2label,
```

```
    label2id=label2id
)

training_args = TrainingArguments(
    output_dir="./content/drive/MyDrive/NoteBook/ner_model",
    eval_strategy="epoch",
    save_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    num_train_epochs=3,
    weight_decay=0.01,
)


trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_datasets["train"],
    eval_dataset=tokenized_datasets["val"],
    processing_class=tokenizer,
    data_collator=DataCollatorForTokenClassification(tokenizer),
    compute_metrics=compute_metrics,
)
```

```
model.safetensors:   0%|          | 0.00/436M [00:00<?, ?B/s]
```

```
Some weights of BertForTokenClassification were not initialized from the model
checkpoint at bert-base-cased 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.
Using the `WANDB_DISABLED` environment variable is deprecated and will be
removed in v5. Use the --report_to flag to control the integrations used for
logging result (for instance --report_to none).
```

0.0.16 Train or Download Model

This cell prompts the user to choose between training the model or downloading a pre-trained model. If the user chooses to train the model, it starts the training process and saves the model to Google Drive. If the user chooses to download the model, it downloads the pre-trained model from Google Drive.

[ ]:
```
choice = input("Do you want to (1) Train the model or (2) Download the␣
 ↪pre-trained model? Enter 1 or 2: ").strip()

if choice == "1":
    print("Training the model...")
```

```
    trainer.train()
    print("Saving the model to Google Drive...")
    model.save_pretrained("/content/drive/MyDrive/NoteBook/ner_model")
    tokenizer.save_pretrained("/content/drive/MyDrive/NoteBook/ner_model")
    print("Model saved to Google Drive.")

    model = AutoModelForTokenClassification.from_pretrained("/content/drive/
 ↪MyDrive/NoteBook/ner_model")
    tokenizer = AutoTokenizer.from_pretrained("/content/drive/MyDrive/NoteBook/
 ↪ner_model")
    print(" model and tokenizer loaded successfully from drive.")

elif choice == "2":
    print("Downloading the pre-trained model...")
    # https://drive.google.com/drive/folders/1-1gu-XgHZ9crDBkdg2TD8e4EvGbGl_OC?
 ↪usp=sharing
    folder_id = "1-1gu-XgHZ9crDBkdg2TD8e4EvGbGl_OC"
    gdown.download_folder(id=folder_id, output="ner_model")
    print("Model downloaded to 'ner_model' folder.")

    model = AutoModelForTokenClassification.from_pretrained("ner_model")
    tokenizer = AutoTokenizer.from_pretrained("ner_model")
    print("Pre-trained model and tokenizer loaded successfully.")
else:
    print("Invalid choice. Please enter 1 or 2.")
```

```
Do you want to (1) Train the model or (2) Download the pre-trained model? Enter
1 or 2: 1
Training the model…
```

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

```
Saving the model to Google Drive…
Model saved to Google Drive.
 model and tokenizer loaded successfully from drive.
```

0.0.17 Load Test Dataset

This cell loads the test dataset from the CSV file and applies the same preprocessing steps as the training and validation datasets.

[ ]:
```
DATA_FILES = {"test": "test.csv"}
test_dataset = load_dataset("csv", data_files=DATA_FILES)
test_dataset = test_dataset.map(convert_str_to_list)
tokenized_test_datasets = test_dataset.map(tokenize_and_align_labels,␣
 ↪batched=True)
```

```
Generating test split: 0 examples [00:00, ? examples/s]
```

```
Map:   0%|          | 0/2852 [00:00<?, ? examples/s]
Map:   0%|          | 0/2852 [00:00<?, ? examples/s]
```

0.0.18 Evaluate Model on Test Dataset

This cell evaluates the model on the test dataset and prints the evaluation results, including accuracy, precision, recall, F1-score, FPR, and FNR.

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

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

{'eval_test_loss': 0.0037880672607570887, 'eval_test_accuracy':
0.9991494914424858, 'eval_test_f1': 0.9963560002470508, 'eval_test_precision':
0.9961714215141411, 'eval_test_recall': 0.9965406473931308, 'eval_test_fpr':
0.0037663799786242405, 'eval_test_fnr': 0.003425495473452389,
'eval_test_runtime': 23.3699, 'eval_test_samples_per_second': 122.037,
'eval_test_steps_per_second': 7.659, 'epoch': 3.0}
```

0.0.19 Initialize NER Pipeline

This cell initializes an NER pipeline using the trained model and tokenizer. The pipeline is used to make predictions on new text data.

[ ]:
```
from transformers import pipeline

ner_pipeline = pipeline("ner", model=model, tokenizer=tokenizer,␣
 ↪aggregation_strategy="simple")

text = "Subhan Rangila and Muhammad Sadiq are data scientists at Google in New␣
 ↪York. Their emails are subhanRangila@gmail.com and muhammad.sadiq@yahoo.com."

predictions = ner_pipeline(text)
for pred in predictions:
    print(f"Entity: {pred['word']}, Label: {pred['entity_group']}, Score:␣
 ↪{pred['score']:.4f}")
```

```
Device set to use cuda:0

Entity: Subhan Rangila, Label: PER, Score: 0.9286
Entity: Muhammad Sadiq, Label: PER, Score: 0.9999
Entity: subhanRangila @ gmail. com, Label: EMAIL, Score: 0.9888
Entity: muhammad, Label: EMAIL, Score: 0.9989
Entity: sadiq @ yahoo. com, Label: EMAIL, Score: 1.0000
```

0.0.20 Redact PII with Pipeline

This cell defines a function to redact PII (e.g., names and emails) from text using the NER pipeline. It replaces PII entities with placeholders like `[NAME]` and `[EMAIL]`.

```
[ ]: def redact_pii_with_pipeline(text, ner_pipeline):
         predictions = ner_pipeline(text)

         predictions = sorted(predictions, key=lambda x: x["start"])

         redacted_text = ""
         prev_end = 0
         for pred in predictions:
             redacted_text += text[prev_end:pred["start"]]

             if pred["entity_group"] == "PER":
                 redacted_text += "[NAME]"
             elif pred["entity_group"] == "EMAIL":
                 redacted_text += "[EMAIL]"
             else:
                 redacted_text += pred["word"]

             prev_end = pred["end"]

         redacted_text += text[prev_end:]

         return redacted_text
```

0.0.21 Test Redaction Function

This cell tests the redaction function on a sample text and prints the original and redacted text.

```
[ ]: text = "Alice Johnson works at Microsoft. Bob Dylan is a researcher at OpenAI.␣
      ↪Their contacts are alice.j@microsoft.com and bobdylan@openai.com."
     redacted_text = redact_pii_with_pipeline(text, ner_pipeline)
     print("Original Text:", text)
     print("Redacted Text:", redacted_text)
```

```
Original Text: Alice Johnson works at Microsoft. Bob Dylan is a researcher at
OpenAI. Their contacts are alice.j@microsoft.com and bobdylan@openai.com.
Redacted Text: [NAME] works at Microsoft. [NAME] is a researcher at OpenAI.
Their contacts are alice.[EMAIL] and [EMAIL].
```

0.0.22 Load Independent Test Dataset

This cell downloads and loads an independent test dataset from Google Drive. The dataset is stored in JSON format.

```
[ ]: # https://drive.google.com/file/d/1E2FjYFDGEeXTwpabkCOaYzZV8aOQqf_h/view?
      ↪usp=sharing
     independent_test_data_file_id = "1E2FjYFDGEeXTwpabkCOaYzZV8aOQqf_h"
```

```
gdown.download(f"https://drive.google.com/uc?
  ↪id={independent_test_data_file_id}", "test_data.json", quiet=False)
```

```
Downloading…
From: https://drive.google.com/uc?id=1E2FjYFDGEeXTwpabkC0aYzZV8aOQqf_h
To: /content/test_data.json
100%|      | 4.19M/4.19M [00:00<00:00, 194MB/s]
```

[ ]:
```
'test_data.json'
```

[ ]:
```
DATA_FILES = {"test_data": "test_data.json"}
test_dataset = load_dataset("json", data_files=DATA_FILES)
tokenized_test_datasets = test_dataset.map(tokenize_and_align_labels,␣
  ↪batched=True)
```

0.0.23 Evaluate Model on Independent Test Dataset

This cell evaluates the model on the independent test dataset and prints the evaluation results. Note that since independent dataset doesn't having email therefore getting fpr, fnr values as nan as can be seen by evaluation

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

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

```
{'eval_test_data_loss': 0.029309110715985298, 'eval_test_data_accuracy':
0.9928402326062473, 'eval_test_data_f1': 0.9389151655429203,
'eval_test_data_precision': 0.9868838586841548, 'eval_test_data_recall':
0.8953934740882917, 'eval_test_data_fpr': nan, 'eval_test_data_fnr': nan,
'eval_test_data_runtime': 30.2626, 'eval_test_data_samples_per_second': 120.611,
'eval_test_data_steps_per_second': 7.567, 'epoch': 3.0}
```

```
<ipython-input-8-426117950397>:32: RuntimeWarning: invalid value encountered in
scalar divide
  fpr[entity_type] = fp / (fp + tp + fn)
<ipython-input-8-426117950397>:33: RuntimeWarning: invalid value encountered in
scalar divide
  fnr[entity_type] = fn / (fn + tp)
```

0.0.24 Load Synthetic Test Dataset

This cell downloads and loads a synthetic test dataset from Google Drive. The dataset is stored in CSV format. Note that when we append synthetic email to the test dataset and then evaluate it we got the values of fnr and fpr

[ ]:
```
# https://drive.google.com/file/d/1-1v9MghJ6XnGDdlKaD4h-se1ZNfk6hYV/view?
  ↪usp=sharing
independent_synthetic_test_data_file_id = "1-1v9MghJ6XnGDdlKaD4h-se1ZNfk6hYV"
```

```
gdown.download(f"https://drive.google.com/uc?
  ↪id={independent_synthetic_test_data_file_id}", "synthetic_test_data.csv",␣
  ↪quiet=False)
```

```
Downloading…
From: https://drive.google.com/uc?id=1-1v9MghJ6XnGDdlKaD4h-se1ZNfk6hYV
To: /content/synthetic_test_data.csv
100%|      | 18.2M/18.2M [00:01<00:00, 11.5MB/s]
```

[ ]:
```
'synthetic_test_data.csv'
```

[ ]:
```
DATA_FILES = {"synthetic_test_data": "synthetic_test_data.csv"}
synthetic_test_dataset = load_dataset("csv", data_files=DATA_FILES)
synthetic_test_dataset = synthetic_test_dataset.map(convert_str_to_list)
tokenized_synthetic_test_dataset = synthetic_test_dataset.
  ↪map(tokenize_and_align_labels, batched=True)
```

```
Generating synthetic_test_data split: 0 examples [00:00, ? examples/s]
```

```
Map:   0%|          | 0/28516 [00:00<?, ? examples/s]
```

```
Map:   0%|          | 0/28516 [00:00<?, ? examples/s]
```

0.0.25 Evaluate Model on Synthetic Test Dataset

This cell evaluates the model on the synthetic test dataset and prints the evaluation results.

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

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

```
{'eval_synthetic_test_data_loss': 0.0007556203636340797,
'eval_synthetic_test_data_accuracy': 0.9998100851803136,
'eval_synthetic_test_data_f1': 0.9990298614892565,
'eval_synthetic_test_data_precision': 0.9989235737351991,
'eval_synthetic_test_data_recall': 0.9991361718642413,
'eval_synthetic_test_data_fpr': 0.0010661004431376582,
'eval_synthetic_test_data_fnr': 0.0008573174886663537,
'eval_synthetic_test_data_runtime': 239.3137,
'eval_synthetic_test_data_samples_per_second': 119.157,
'eval_synthetic_test_data_steps_per_second': 7.45, 'epoch': 3.0}
```

[ ]:

[ ]: