Data_Preparation

April 21, 2025

1 PII Masking : Data Preparation and Analysis

1.1 Objective

This notebook prepares and analyzes a dataset for training a model to detect and mask Personally Identifiable Information (PII), specifically names and email addresses. The dataset used is a subset of the WikiNeural dataset, which is enriched with synthetic email addresses if they are not already present.

1.2 Steps

1. **Data Loading**: Load the dataset from Google Drive.
2. **Data Inspection**: Analyze the dataset to understand its structure, NER tag distribution, and sequence lengths.
3. **Synthetic Email Generation**: Inject synthetic email addresses into the dataset to ensure a realistic distribution of PII.
4. **Data Splitting**: Split the dataset into training, validation, and test sets.
5. **Data Saving**: Save the processed datasets for future use.

[1]:

```
!pip install datasets
!pip install -U accelerate
!pip install -U transformers
!pip install seqeval
!pip install gdown
!pip install pandas scikit-learn matplotlib seaborn
```

```
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
8.0 MB/s eta 0:00:00
Downloading dill-0.3.8-py3-none-any.whl (116 kB)
                         116.3/116.3 kB
6.0 MB/s eta 0:00:00
Downloading multiprocess-0.70.16-py311-none-any.whl (143 kB)
                         143.5/143.5 kB
7.8 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
11.1 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
8.3 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
4.1 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
44.8 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
31.5 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
38.5 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
4.6 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
11.9 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
7.4 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
6.7 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
81.1 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
2.2 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
51.4 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
1.6 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=048163487699b08d3126c897091e59400c30e9fdeefc76d526570178242750b0
  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: matplotlib in /usr/local/lib/python3.11/dist-
packages (3.10.0)
Requirement already satisfied: seaborn in /usr/local/lib/python3.11/dist-
packages (0.13.2)
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: contourpy>=1.0.1 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (1.3.1)
```

```
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-
packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (4.56.0)
Requirement already satisfied: kiwisolver>=1.3.1 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (1.4.8)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (24.2)
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.11/dist-
packages (from matplotlib) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in
/usr/local/lib/python3.11/dist-packages (from matplotlib) (3.2.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
```

1.2.1 Import Required Libraries

This cell imports the necessary Python libraries for data manipulation, analysis, and visualization. - `pandas`: For data manipulation. - `train_test_split`: For splitting the dataset. - `random`: For generating synthetic email addresses. - `gdown`: For downloading files from Google Drive. - `ast`: For safely evaluating strings as Python expressions.

[2]:
```python
import pandas as pd
from sklearn.model_selection import train_test_split
import random
import gdown
import ast
```

1.2.2 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. The commented path are for the local storage. Adjust them according to your need

[3]:
```python
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call
drive.mount("/content/drive", force_remount=True).
```

1.2.3 Download Dataset from Google Drive

This cell downloads the dataset (`data.json`) from Google Drive using the `gdown` library. The dataset is then loaded into a Pandas DataFrame for further processing.

[4]:

```
# https://drive.google.com/file/d/1pmYdpJHdaYuPnG1PvPGND-7KABtRHPi9/view?
 ↪usp=sharing
data_file_id ="1pmYdpJHdaYuPnG1PvPGND-7KABtRHPi9"

gdown.download(f"https://drive.google.com/uc?id={data_file_id}", "data.json",␣
 ↪quiet=False)
data = pd.read_json('data.json')
print(data.head())
```

```
Downloading…
From: https://drive.google.com/uc?id=1pmYdpJHdaYuPnG1PvPGND-7KABtRHPi9
To: /content/data.json
100%|      | 34.0M/34.0M [00:00<00:00, 54.8MB/s]

  lang                                           ner_tags  \
0   en  [O, O, O, O, B-PER, I-PER, O, O, O, O, B-PER, …
1   en  [O, O, O, O, B-PER, I-PER, O, O, O, O, O, O, O…
2   en                 [O, O, O, O, O, O, O, O, B-PER, O]
3   en  [O, O, O, B-PER, O, O, O, O, O, O, O, O, O, O,…
4   en  [O, O, O, O, B-PER, I-PER, I-PER, O, O, O, O, …

                                            sequence  \
0  Since then , only Terry Bradshaw in 147 games …
1  He was portrayed by Anthony Perkins in the 196…
2  The egg eventually hatches , revealing a baby …
3  In the video Kelis is walking down a street in…
4  According to food writer Sharon Tyler Herbst ,…

                                              tokens
0  [Since, then, ,, only, Terry, Bradshaw, in, 14…
1  [He, was, portrayed, by, Anthony, Perkins, in,…
2  [The, egg, eventually, hatches, ,, revealing, …
3  [In, the, video, Kelis, is, walking, down, a, …
4  [According, to, food, writer, Sharon, Tyler, H…
```

1.2.4 Dataset Inspection and Analysis

This section focuses on inspecting and analyzing the dataset to understand its structure, content, and distribution of Named Entity Recognition (NER) tags. The following steps are performed:

1. **Dataset Overview**:
   - The first few rows of the dataset are displayed to provide a snapshot of its structure.
   - The columns and their data types are listed to understand the schema of the dataset.
   - Missing values are checked to ensure data quality.
2. **NER Tag Analysis**:
   - The frequency of each NER tag is counted to understand the distribution of entity types (e.g., names, email addresses).
   - The most common and rare NER tags are identified, which helps in understanding the dataset's balance and potential biases.

3. **Token Analysis**:
    - Sample tokens from the dataset are displayed to provide insights into the tokenized text data.
4. **Visualization**:
    - A bar plot is created to visualize the distribution of NER tags, highlighting the most common and rare tags.
    - A histogram is generated to visualize the distribution of

[ ]:
```python
print("Columns in the dataset:", data.columns)
print("\nData types:\n", data.dtypes)
```

```
Columns in the dataset: Index(['lang', 'ner_tags', 'sequence', 'tokens'],
dtype='object')

Data types:
 lang        object
ner_tags    object
sequence    object
tokens      object
dtype: object
```

[ ]:
```python
from collections import Counter

all_ner_tags = [tag for sublist in data["ner_tags"] for tag in sublist]

ner_tag_counts = Counter(all_ner_tags)
print("Frequency of NER tags:")
print(ner_tag_counts)
```

```
Frequency of NER tags:
Counter({'O': 639055, 'B-PER': 40264, 'I-PER': 29466})
```

[ ]:
```python
print("Sample tokens:")
print(data["tokens"].head())
```

```
Sample tokens:
0    [Since, then, ,, only, Terry, Bradshaw, in, 14…
1    [He, was, portrayed, by, Anthony, Perkins, in,…
2    [The, egg, eventually, hatches, ,, revealing, …
3    [In, the, video, Kelis, is, walking, down, a, …
4    [According, to, food, writer, Sharon, Tyler, H…
Name: tokens, dtype: object
```

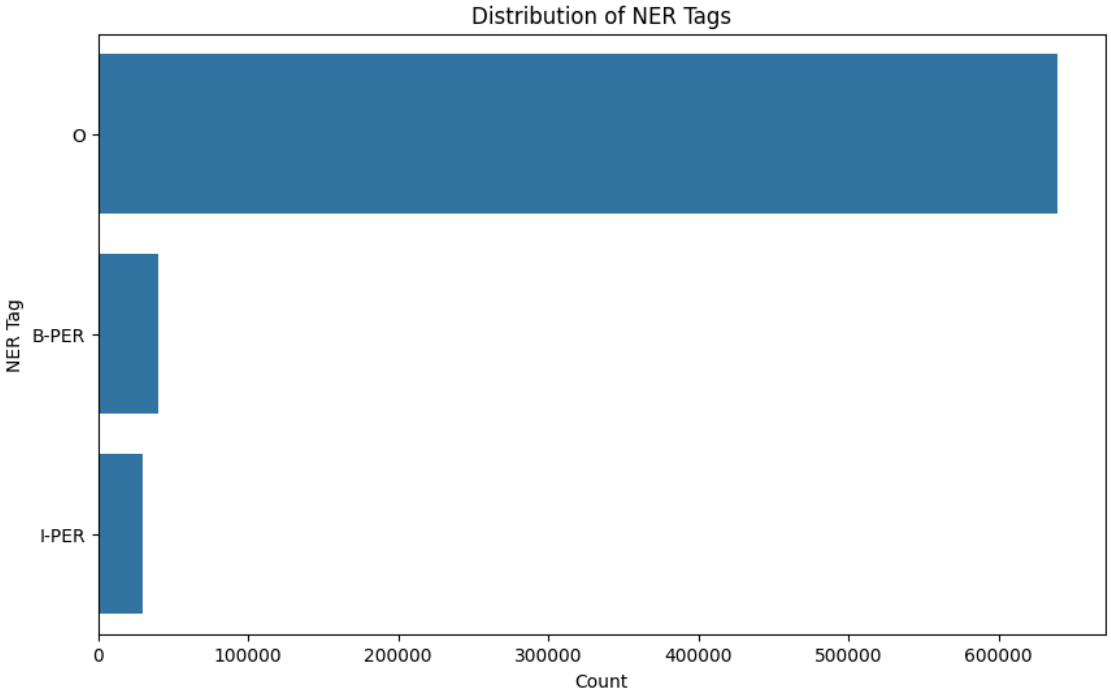
[ ]:
```python
print("Missing values in each column:")
print(data.isnull().sum())
```

```
Missing values in each column:
lang        0
```

```
ner_tags    0
sequence    0
tokens      0
dtype: int64
```

```
[ ]: import matplotlib.pyplot as plt
     import seaborn as sns

     plt.figure(figsize=(10, 6))
     sns.countplot(y=all_ner_tags, order=[tag for tag, _ in ner_tag_counts.
       ↪most_common()])
     plt.title("Distribution of NER Tags")
     plt.xlabel("Count")
     plt.ylabel("NER Tag")
     plt.show()
```



1.2.5 Generate Synthetic Emails

This cell defines a function to generate synthetic email addresses using a combination of names and random domains. The function ensures that the generated emails are realistic.

```
[5]: def generate_email(name):
         domains = ["gmail.com", "yahoo.com", "outlook.com", "fastnu.edu.pk"]
         formatted_name = name.lower().replace(" ", ".")
         return f"{formatted_name}{random.randint(10, 99)}@{random.choice(domains)}"
```

1.2.6 Apply Email Injection

This cell applies the `append_emails` function to the entire dataset, enriching it with synthetic email addresses. The modified dataset is then saved to a CSV file for future use.

[6]:
```
def append_emails(data, email_probability=0.05):
    sequence = data["sequence"]
    tokens = data["tokens"]
    ner_tags = data["ner_tags"]

    for i, token in enumerate(tokens):
        if ner_tags[i] == "B-PER":
            name = token.lower()

            if i + 1 < len(tokens) and ner_tags[i + 1] == "I-PER":
                name += "." + tokens[i + 1].lower()
                i += 1

            email = generate_email(name)
            email_tokens = email.split("@")[0].split(".") + ["@", email.
 ↪split("@")[1]]

            tokens.extend(email_tokens)
            ner_tags.extend(["B-EMAIL"] + ["I-EMAIL"] * (len(email_tokens) - 1))

            sequence += f" {email}"

    data["sequence"] = sequence
    data["tokens"] = tokens
    data["ner_tags"] = ner_tags
    return data
```

[7]:
```
data = data.apply(append_emails, axis=1)
data.to_csv("/content/drive/MyDrive/NoteBook/synthetic_data.csv", index=False)
data.to_csv("synthetic_data.csv", index=False)
```

1.2.7 Load Modified Dataset

This cell loads the modified dataset (with synthetic emails) from the CSV file and displays the first few rows to verify the changes.

[8]:
```
data = pd.read_csv("/content/drive/MyDrive/NoteBook/synthetic_data.csv")
# data = pd.read_csv("synthetic_data.csv")
print(data.head()['tokens'][0])
```

```
['Since', 'then', ',', 'only', 'Terry', 'Bradshaw', 'in', '147', 'games', ',',
'Joe', 'Montana', 'in', '139', 'games', ',', 'and', 'Tom', 'Brady', 'in', '131',
'games', 'have', 'reached', '100', 'wins', 'more', 'quickly', '.', 'terry',
'bradshaw25', '@', 'yahoo.com', 'joe', 'montana28', '@', 'outlook.com', 'tom',
'brady22', '@', 'yahoo.com']
```

1.2.8 Save Dataset Splits

This cell saves the training, validation, and test sets to CSV files. These files will be used for model training and evaluation.

[ ]:

```
train_data, temp_data = train_test_split(data, test_size=0.2, random_state=42)
val_data, test_data = train_test_split(temp_data, test_size=0.5,␣
 ↪random_state=42)

train_data.to_csv("/content/drive/MyDrive/NoteBook/train.csv", index=False)
val_data.to_csv("/content/drive/MyDrive/NoteBook/val.csv", index=False)
test_data.to_csv("/content/drive/MyDrive/NoteBook/test.csv", index=False)
# train_data.to_csv("train.csv", index=False)
# val_data.to_csv("val.csv", index=False)
# test_data.to_csv("test.csv", index=False)
```

1.2.9 Download and Prepare Test Data

This cell downloads a test dataset from Google Drive using `gdown`. The dataset is saved as a JSON file (`test_data.json`). After downloading, the dataset is loaded into a Pandas DataFrame. A function (`append_emails`) is applied to the dataset to add or modify email-related data. Finally, the processed dataset is saved as a CSV file (`synthetic_test_data.csv`) for further use. The first few rows of the dataset are printed to verify the data.

[ ]:

```
# https://drive.google.com/file/d/1pmYdpJHdaYuPnG1PvPGND-7KABtRHPi9/view?
 ↪usp=sharing
test_data_file_id ="1E2FjYFDGEeXTwpabkC0aYzZV8a0Qqf_h"

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

```
Downloading…
From: https://drive.google.com/uc?id=1E2FjYFDGEeXTwpabkC0aYzZV8a0Qqf_h
To: /content/test_data.json
100%|     | 4.19M/4.19M [00:00<00:00, 95.8MB/s]
```

[ ]:

```
test_data = test_data.apply(append_emails, axis=1)
data.to_csv("/content/drive/MyDrive/NoteBook/synthetic_test_data.csv",␣
 ↪index=False)
# data.to_csv("synthetic_test_data.csv", index=False)
```

[ ]:

```
print(test_data.head())
```

```
  lang                                            ner_tags  \
0   en  [O, O, O, O, O, O, O, O, O, B-PER, I-PER, O, B…
1   en  [O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, …
2   en  [O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, …
3   en  [O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, …
4   en  [O, O, O, O, O, O, O, B-PER, I-PER, O, B-EMAIL…

                                          sequence  \
0  included future Rage Against the Machine and A…
1  The city voted 53.5 percent in favor of the ma…
2  It was not until about 1907 - 1909 that he pro…
3  Always in his bowler hat , he was a witty pres…
4  The music was composed and conducted by Ronald…

                                            tokens
0  [included, future, Rage, Against, the, Machine…
1  [The, city, voted, 53.5, percent, in, favor, o…
2  [It, was, not, until, about, 1907, -, 1909, th…
3  [Always, in, his, bowler, hat, ,, he, was, a, …
4  [The, music, was, composed, and, conducted, by…
```

[ ]: