models_tutorial_classifier_2025

March 19, 2025

0.1 README

This is an example tutorial on how to use pre-trained models to predict occupation codes for online advertisements.

Note that the software allows you to use these models to fine tune them for your training and test data.

Models: https://repod.icm.edu.pl/dataset.xhtml?persistentId=doi:10.18150/OCUTSI

Publication: Beręsewicz, M., Wydmuch, M., Cherniaiev, H., & Pater, R. (2024). Multilingual hierarchical classification of job advertisements for job vacancy statistics. arXiv preprint arXiv:2411.03779.

How to cite:

```
@misc{beręsewicz2024multilingualhierarchicalclassificationjob,
      title={Multilingual hierarchical classification of job advertisements for job vacancy sta
      author={Maciej Beręsewicz and Marek Wydmuch and Herman Cherniaiev and Robert Pater},
      year={2024},
      eprint={2411.03779},
      archivePrefix={arXiv},
      primaryClass={stat.AP},
      url={https://arxiv.org/abs/2411.03779},
}

@data{OCUTSI_2024,
      author = {Beręsewicz, Maciej and Wydmuch, Marek and Pater, Robert and Cherniaiev, Herman
      publisher = {RepOD},
      title = "{Job offers classifiers for ISCO and KZiS 2023}",
      year = {2024},
      version = {V1},
      doi = {10.18150/OCUTSI},
      url = {https://doi.org/10.18150/OCUTSI}
}
```

0.2 Setup (installation and models)

Note that we check the code for the Python versions: 3.9, 3.10 and 3.11.

```
[ ]: !python --version
```

```
Python 3.11.11
```

[ ]:

```
!rm -rf job_offers_classifier
```

[ ]:

```
!git clone https://github.com/OJALAB/job-ads-classifier.git
!mv job-ads-classifier/* ./
!rm -rf job-ads-classifier
```

```
Cloning into 'job-ads-classifier'…
remote: Enumerating objects: 30, done.
remote: Counting objects: 100% (30/30), done.
remote: Compressing objects: 100% (25/25), done.
remote: Total 30 (delta 5), reused 23 (delta 3), pack-reused 0 (from 0)
Receiving objects: 100% (30/30), 2.14 MiB | 13.47 MiB/s, done.
Resolving deltas: 100% (5/5), done.
```

[ ]:

```
!pip install virtualenv
!virtualenv classifier
!classifier/bin/pip install --upgrade pip
!classifier/bin/pip install torch
!classifier/bin/pip install -r requirements-colab.txt
```

```
Collecting virtualenv
  Downloading virtualenv-20.29.3-py3-none-any.whl.metadata (4.5 kB)
Collecting distlib<1,>=0.3.7 (from virtualenv)
  Downloading distlib-0.3.9-py2.py3-none-any.whl.metadata (5.2 kB)
Requirement already satisfied: filelock<4,>=3.12.2 in
/usr/local/lib/python3.11/dist-packages (from virtualenv) (3.17.0)
Requirement already satisfied: platformdirs<5,>=3.9.1 in
/usr/local/lib/python3.11/dist-packages (from virtualenv) (4.3.6)
Downloading virtualenv-20.29.3-py3-none-any.whl (4.3 MB)
                         4.3/4.3 MB
78.5 MB/s eta 0:00:00
Downloading distlib-0.3.9-py2.py3-none-any.whl (468 kB)
                         469.0/469.0 kB
30.2 MB/s eta 0:00:00
Installing collected packages: distlib, virtualenv
Successfully installed distlib-0.3.9 virtualenv-20.29.3
created virtual environment CPython3.11.11.final.0-64 in 1572ms
  creator CPython3Posix(dest=/content/classifier, clear=False,
no_vcs_ignore=False, global=False)
  seeder FromAppData(download=False, pip=bundle, setuptools=bundle,
wheel=bundle, via=copy, app_data_dir=/root/.local/share/virtualenv)
    added seed packages: pip==25.0.1, setuptools==75.8.0, wheel==0.45.1
  activators BashActivator,CShellActivator,FishActivator,NushellActivator,PowerS
hellActivator,PythonActivator
Requirement already satisfied: pip in ./classifier/lib/python3.11/site-packages
(25.0.1)
Collecting torch
```

```
  Downloading torch-2.6.0-cp311-cp311-manylinux1_x86_64.whl.metadata (28 kB)
Collecting filelock (from torch)
  Downloading filelock-3.18.0-py3-none-any.whl.metadata (2.9 kB)
Collecting typing-extensions>=4.10.0 (from torch)
  Downloading typing_extensions-4.12.2-py3-none-any.whl.metadata (3.0 kB)
Collecting networkx (from torch)
  Downloading networkx-3.4.2-py3-none-any.whl.metadata (6.3 kB)
Collecting jinja2 (from torch)
  Downloading jinja2-3.1.6-py3-none-any.whl.metadata (2.9 kB)
Collecting fsspec (from torch)
  Downloading fsspec-2025.3.0-py3-none-any.whl.metadata (11 kB)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch)
  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)
  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)
  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)
  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)
  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)
  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)
  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)
  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)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-
manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparselt-cu12==0.6.2 (from torch)
  Downloading nvidia_cusparselt_cu12-0.6.2-py3-none-
manylinux2014_x86_64.whl.metadata (6.8 kB)
Collecting nvidia-nccl-cu12==2.21.5 (from torch)
  Downloading nvidia_nccl_cu12-2.21.5-py3-none-manylinux2014_x86_64.whl.metadata
(1.8 kB)
Collecting nvidia-nvtx-cu12==12.4.127 (from torch)
  Downloading nvidia_nvtx_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.7 kB)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch)
```

```
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting triton==3.2.0 (from torch)
  Downloading
triton-3.2.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
(1.4 kB)
Collecting sympy==1.13.1 (from torch)
  Downloading sympy-1.13.1-py3-none-any.whl.metadata (12 kB)
Collecting mpmath<1.4,>=1.1.0 (from sympy==1.13.1->torch)
  Downloading mpmath-1.3.0-py3-none-any.whl.metadata (8.6 kB)
Collecting MarkupSafe>=2.0 (from jinja2->torch)
  Downloading MarkupSafe-3.0.2-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (4.0 kB)
Downloading torch-2.6.0-cp311-cp311-manylinux1_x86_64.whl (766.7 MB)
                         766.7/766.7 MB
29.6 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
59.0 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
109.4 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
110.1 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
48.4 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
40.7 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
65.6 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
79.2 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
65.9 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
73.0 MB/s eta 0:00:00
Downloading nvidia_cusparselt_cu12-0.6.2-py3-none-manylinux2014_x86_64.whl
(150.1 MB)
                         150.1/150.1 MB
68.3 MB/s eta 0:00:00
Downloading nvidia_nccl_cu12-2.21.5-py3-none-manylinux2014_x86_64.whl
(188.7 MB)
                         188.7/188.7 MB
66.0 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
110.1 MB/s eta 0:00:00
Downloading nvidia_nvtx_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl
(99 kB)
Downloading sympy-1.13.1-py3-none-any.whl (6.2 MB)
                         6.2/6.2 MB
104.6 MB/s eta 0:00:00
Downloading
triton-3.2.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (253.2
MB)
                         253.2/253.2 MB
54.0 MB/s eta 0:00:00
Downloading typing_extensions-4.12.2-py3-none-any.whl (37 kB)
Downloading filelock-3.18.0-py3-none-any.whl (16 kB)
Downloading fsspec-2025.3.0-py3-none-any.whl (193 kB)
Downloading jinja2-3.1.6-py3-none-any.whl (134 kB)
Downloading networkx-3.4.2-py3-none-any.whl (1.7 MB)
                         1.7/1.7 MB
81.4 MB/s eta 0:00:00
Downloading
MarkupSafe-3.0.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (23
kB)
Downloading mpmath-1.3.0-py3-none-any.whl (536 kB)
                         536.2/536.2 kB
27.8 MB/s eta 0:00:00
Installing collected packages: triton, nvidia-cusparselt-cu12, mpmath,
typing-extensions, sympy, nvidia-nvtx-cu12, nvidia-nvjitlink-cu12, nvidia-nccl-
cu12, nvidia-curand-cu12, nvidia-cufft-cu12, nvidia-cuda-runtime-cu12, nvidia-
cuda-nvrtc-cu12, nvidia-cuda-cupti-cu12, nvidia-cublas-cu12, networkx,
MarkupSafe, fsspec, filelock, nvidia-cusparse-cu12, nvidia-cudnn-cu12, jinja2,
nvidia-cusolver-cu12, torch
Successfully installed MarkupSafe-3.0.2 filelock-3.18.0 fsspec-2025.3.0
jinja2-3.1.6 mpmath-1.3.0 networkx-3.4.2 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-cusparselt-cu12-0.6.2 nvidia-nccl-cu12-2.21.5 nvidia-
nvjitlink-cu12-12.4.127 nvidia-nvtx-cu12-12.4.127 sympy-1.13.1 torch-2.6.0
triton-3.2.0 typing-extensions-4.12.2
Collecting click==8.0.3 (from -r requirements-colab.txt (line 1))
  Downloading click-8.0.3-py3-none-any.whl.metadata (3.2 kB)
Collecting jsbeautifier==1.15.1 (from -r requirements-colab.txt (line 2))
  Downloading jsbeautifier-1.15.1.tar.gz (75 kB)
  Installing build dependencies … done
  Getting requirements to build wheel … done
  Preparing metadata (pyproject.toml) … done
Collecting lxml==5.3.0 (from -r requirements-colab.txt (line 3))
  Downloading lxml-5.3.0-cp311-cp311-manylinux_2_28_x86_64.whl.metadata (3.8 kB)
Collecting napkinxc==0.7.1 (from -r requirements-colab.txt (line 4))
  Downloading napkinxc-0.7.1-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (6.0 kB)
Collecting numpy==1.24.3 (from -r requirements-colab.txt (line 5))
  Downloading
numpy-1.24.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
(5.6 kB)
Collecting pandas==2.2.2 (from -r requirements-colab.txt (line 6))
  Downloading
pandas-2.2.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
(19 kB)
Collecting pyreadr==0.5.2 (from -r requirements-colab.txt (line 7))
  Downloading pyreadr-0.5.2-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (1.2 kB)
Collecting pystempel==1.2.0 (from -r requirements-colab.txt (line 8))
  Downloading pystempel-1.2.0-py3-none-any.whl.metadata (7.9 kB)
Collecting pytorch_lightning>=2.0.0 (from -r requirements-colab.txt (line 9))
  Downloading pytorch_lightning-2.5.0.post0-py3-none-any.whl.metadata (21 kB)
Collecting sacremoses==0.1.1 (from -r requirements-colab.txt (line 10))
  Downloading sacremoses-0.1.1-py3-none-any.whl.metadata (8.3 kB)
Collecting scikit_learn==1.2.2 (from -r requirements-colab.txt (line 11))
  Downloading scikit_learn-1.2.2-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (11 kB)
Collecting scipy>=1.11.0 (from -r requirements-colab.txt (line 12))
  Downloading
scipy-1.15.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
(61 kB)
Collecting stop_words==2018.7.23 (from -r requirements-colab.txt (line 13))
  Downloading stop-words-2018.7.23.tar.gz (31 kB)
  Preparing metadata (setup.py) … done
Collecting torchmetrics==1.4.1 (from -r requirements-colab.txt (line 14))
  Downloading torchmetrics-1.4.1-py3-none-any.whl.metadata (20 kB)
Collecting tqdm==4.66.5 (from -r requirements-colab.txt (line 15))
  Downloading tqdm-4.66.5-py3-none-any.whl.metadata (57 kB)
```

```
Collecting transformers==4.44.2 (from -r requirements-colab.txt (line 16))
  Downloading transformers-4.44.2-py3-none-any.whl.metadata (43 kB)
Collecting six>=1.13.0 (from jsbeautifier==1.15.1->-r requirements-colab.txt
(line 2))
  Downloading six-1.17.0-py2.py3-none-any.whl.metadata (1.7 kB)
Collecting editorconfig>=0.12.2 (from jsbeautifier==1.15.1->-r requirements-
colab.txt (line 2))
  Downloading EditorConfig-0.17.0-py3-none-any.whl.metadata (3.8 kB)
Collecting gdown (from napkinxc==0.7.1->-r requirements-colab.txt (line 4))
  Downloading gdown-5.2.0-py3-none-any.whl.metadata (5.8 kB)
Collecting python-dateutil>=2.8.2 (from pandas==2.2.2->-r requirements-colab.txt
(line 6))
  Downloading python_dateutil-2.9.0.post0-py2.py3-none-any.whl.metadata (8.4 kB)
Collecting pytz>=2020.1 (from pandas==2.2.2->-r requirements-colab.txt (line 6))
  Downloading pytz-2025.1-py2.py3-none-any.whl.metadata (22 kB)
Collecting tzdata>=2022.7 (from pandas==2.2.2->-r requirements-colab.txt (line
6))
  Downloading tzdata-2025.1-py2.py3-none-any.whl.metadata (1.4 kB)
Collecting sortedcontainers (from pystempel==1.2.0->-r requirements-colab.txt
(line 8))
  Downloading sortedcontainers-2.4.0-py2.py3-none-any.whl.metadata (10 kB)
Collecting regex (from sacremoses==0.1.1->-r requirements-colab.txt (line 10))
  Downloading regex-2024.11.6-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (40 kB)
Collecting joblib (from sacremoses==0.1.1->-r requirements-colab.txt (line 10))
  Downloading joblib-1.4.2-py3-none-any.whl.metadata (5.4 kB)
Collecting threadpoolctl>=2.0.0 (from scikit_learn==1.2.2->-r requirements-
colab.txt (line 11))
  Downloading threadpoolctl-3.6.0-py3-none-any.whl.metadata (13 kB)
Collecting packaging>17.1 (from torchmetrics==1.4.1->-r requirements-colab.txt
(line 14))
  Downloading packaging-24.2-py3-none-any.whl.metadata (3.2 kB)
Requirement already satisfied: torch>=1.10.0 in
./classifier/lib/python3.11/site-packages (from torchmetrics==1.4.1->-r
requirements-colab.txt (line 14)) (2.6.0)
Collecting lightning-utilities>=0.8.0 (from torchmetrics==1.4.1->-r
requirements-colab.txt (line 14))
  Downloading lightning_utilities-0.14.1-py3-none-any.whl.metadata (5.6 kB)
Requirement already satisfied: filelock in ./classifier/lib/python3.11/site-
packages (from transformers==4.44.2->-r requirements-colab.txt (line 16))
(3.18.0)
Collecting huggingface-hub<1.0,>=0.23.2 (from transformers==4.44.2->-r
requirements-colab.txt (line 16))
  Downloading huggingface_hub-0.29.3-py3-none-any.whl.metadata (13 kB)
Collecting pyyaml>=5.1 (from transformers==4.44.2->-r requirements-colab.txt
(line 16))
  Downloading
PyYAML-6.0.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
```

```
(2.1 kB)
Collecting requests (from transformers==4.44.2->-r requirements-colab.txt (line
16))
  Downloading requests-2.32.3-py3-none-any.whl.metadata (4.6 kB)
Collecting safetensors>=0.4.1 (from transformers==4.44.2->-r requirements-
colab.txt (line 16))
  Downloading safetensors-0.5.3-cp38-abi3-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (3.8 kB)
Collecting tokenizers<0.20,>=0.19 (from transformers==4.44.2->-r requirements-
colab.txt (line 16))
  Downloading tokenizers-0.19.1-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (6.7 kB)
Requirement already satisfied: fsspec>=2022.5.0 in
./classifier/lib/python3.11/site-packages (from
fsspec[http]>=2022.5.0->pytorch_lightning>=2.0.0->-r requirements-colab.txt
(line 9)) (2025.3.0)
Requirement already satisfied: typing-extensions>=4.4.0 in
./classifier/lib/python3.11/site-packages (from pytorch_lightning>=2.0.0->-r
requirements-colab.txt (line 9)) (4.12.2)
Collecting aiohttp!=4.0.0a0,!=4.0.0a1 (from
fsspec[http]>=2022.5.0->pytorch_lightning>=2.0.0->-r requirements-colab.txt
(line 9))
  Downloading aiohttp-3.11.14-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (7.7 kB)
Requirement already satisfied: setuptools in ./classifier/lib/python3.11/site-
packages (from lightning-utilities>=0.8.0->torchmetrics==1.4.1->-r requirements-
colab.txt (line 14)) (75.8.0)
Requirement already satisfied: networkx in ./classifier/lib/python3.11/site-
packages (from torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt
(line 14)) (3.4.2)
Requirement already satisfied: jinja2 in ./classifier/lib/python3.11/site-
packages (from torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt
(line 14)) (3.1.6)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(12.4.127)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(12.4.127)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(12.4.127)
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
```

```
(9.1.0.70)
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(12.4.5.8)
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(11.2.1.3)
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(10.3.5.147)
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(11.6.1.9)
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(12.3.1.170)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14)) (0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(12.4.127)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(12.4.127)
Requirement already satisfied: triton==3.2.0 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14)) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in
./classifier/lib/python3.11/site-packages (from
torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
./classifier/lib/python3.11/site-packages (from
sympy==1.13.1->torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt
(line 14)) (1.3.0)
Collecting beautifulsoup4 (from gdown->napkinxc==0.7.1->-r requirements-
colab.txt (line 4))
  Downloading beautifulsoup4-4.13.3-py3-none-any.whl.metadata (3.8 kB)
Collecting charset-normalizer<4,>=2 (from requests->transformers==4.44.2->-r
requirements-colab.txt (line 16))
  Downloading charset_normalizer-3.4.1-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (35 kB)
Collecting idna<4,>=2.5 (from requests->transformers==4.44.2->-r requirements-
colab.txt (line 16))
  Downloading idna-3.10-py3-none-any.whl.metadata (10 kB)
Collecting urllib3<3,>=1.21.1 (from requests->transformers==4.44.2->-r
requirements-colab.txt (line 16))
  Downloading urllib3-2.3.0-py3-none-any.whl.metadata (6.5 kB)
Collecting certifi>=2017.4.17 (from requests->transformers==4.44.2->-r
requirements-colab.txt (line 16))
  Downloading certifi-2025.1.31-py3-none-any.whl.metadata (2.5 kB)
Collecting aiohappyeyeballs>=2.3.0 (from
aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]>=2022.5.0->pytorch_lightning>=2.0.0->-r
requirements-colab.txt (line 9))
  Downloading aiohappyeyeballs-2.6.1-py3-none-any.whl.metadata (5.9 kB)
Collecting aiosignal>=1.1.2 (from
aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]>=2022.5.0->pytorch_lightning>=2.0.0->-r
requirements-colab.txt (line 9))
  Downloading aiosignal-1.3.2-py2.py3-none-any.whl.metadata (3.8 kB)
Collecting attrs>=17.3.0 (from
aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]>=2022.5.0->pytorch_lightning>=2.0.0->-r
requirements-colab.txt (line 9))
  Downloading attrs-25.3.0-py3-none-any.whl.metadata (10 kB)
Collecting frozenlist>=1.1.1 (from
aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]>=2022.5.0->pytorch_lightning>=2.0.0->-r
requirements-colab.txt (line 9))
  Downloading frozenlist-1.5.0-cp311-cp311-
manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_6
4.whl.metadata (13 kB)
Collecting multidict<7.0,>=4.5 (from
aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]>=2022.5.0->pytorch_lightning>=2.0.0->-r
requirements-colab.txt (line 9))
  Downloading multidict-6.2.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (4.9 kB)
Collecting propcache>=0.2.0 (from
aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]>=2022.5.0->pytorch_lightning>=2.0.0->-r
requirements-colab.txt (line 9))
  Downloading propcache-0.3.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (10 kB)
Collecting yarl<2.0,>=1.17.0 (from
aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]>=2022.5.0->pytorch_lightning>=2.0.0->-r
requirements-colab.txt (line 9))
  Downloading
yarl-1.18.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
```

```
(69 kB)
Collecting soupsieve>1.2 (from beautifulsoup4->gdown->napkinxc==0.7.1->-r
requirements-colab.txt (line 4))
  Downloading soupsieve-2.6-py3-none-any.whl.metadata (4.6 kB)
Requirement already satisfied: MarkupSafe>=2.0 in
./classifier/lib/python3.11/site-packages (from
jinja2->torch>=1.10.0->torchmetrics==1.4.1->-r requirements-colab.txt (line 14))
(3.0.2)
Collecting PySocks!=1.5.7,>=1.5.6 (from
requests[socks]->gdown->napkinxc==0.7.1->-r requirements-colab.txt (line 4))
  Downloading PySocks-1.7.1-py3-none-any.whl.metadata (13 kB)
Downloading click-8.0.3-py3-none-any.whl (97 kB)
Downloading lxml-5.3.0-cp311-cp311-manylinux_2_28_x86_64.whl (5.0 MB)
                         5.0/5.0 MB
118.8 MB/s eta 0:00:00
Downloading
napkinxc-0.7.1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (555
kB)
                         555.1/555.1 kB
28.0 MB/s eta 0:00:00
Downloading
numpy-1.24.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (17.3
MB)
                         17.3/17.3 MB
122.1 MB/s eta 0:00:00
Downloading
pandas-2.2.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (13.0
MB)
                         13.0/13.0 MB
125.5 MB/s eta 0:00:00
Downloading
pyreadr-0.5.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (416
kB)
Downloading pystempel-1.2.0-py3-none-any.whl (2.7 MB)
                         2.7/2.7 MB
9.2 MB/s eta 0:00:00
Downloading sacremoses-0.1.1-py3-none-any.whl (897 kB)
                         897.5/897.5 kB
46.6 MB/s eta 0:00:00
Downloading
scikit_learn-1.2.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(9.6 MB)
                         9.6/9.6 MB
133.9 MB/s eta 0:00:00
Downloading torchmetrics-1.4.1-py3-none-any.whl (866 kB)
                         866.2/866.2 kB
43.5 MB/s eta 0:00:00
Downloading tqdm-4.66.5-py3-none-any.whl (78 kB)
Downloading transformers-4.44.2-py3-none-any.whl (9.5 MB)
                         9.5/9.5 MB
133.2 MB/s eta 0:00:00
Downloading pytorch_lightning-2.5.0.post0-py3-none-any.whl (819 kB)
                         819.3/819.3 kB
43.4 MB/s eta 0:00:00
Downloading
scipy-1.15.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (37.6
MB)
                         37.6/37.6 MB
51.2 MB/s eta 0:00:00
Downloading EditorConfig-0.17.0-py3-none-any.whl (16 kB)
Downloading huggingface_hub-0.29.3-py3-none-any.whl (468 kB)
Downloading joblib-1.4.2-py3-none-any.whl (301 kB)
Downloading lightning_utilities-0.14.1-py3-none-any.whl (28 kB)
Downloading packaging-24.2-py3-none-any.whl (65 kB)
Downloading python_dateutil-2.9.0.post0-py2.py3-none-any.whl (229 kB)
Downloading pytz-2025.1-py2.py3-none-any.whl (507 kB)
Downloading
PyYAML-6.0.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (762 kB)
                         763.0/763.0 kB
41.8 MB/s eta 0:00:00
Downloading
regex-2024.11.6-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (792
kB)
                         792.7/792.7 kB
47.3 MB/s eta 0:00:00
Downloading
safetensors-0.5.3-cp38-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (471
kB)
Downloading six-1.17.0-py2.py3-none-any.whl (11 kB)
Downloading threadpoolctl-3.6.0-py3-none-any.whl (18 kB)
Downloading
tokenizers-0.19.1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(3.6 MB)
                         3.6/3.6 MB
126.7 MB/s eta 0:00:00
Downloading tzdata-2025.1-py2.py3-none-any.whl (346 kB)
Downloading gdown-5.2.0-py3-none-any.whl (18 kB)
Downloading requests-2.32.3-py3-none-any.whl (64 kB)
Downloading sortedcontainers-2.4.0-py2.py3-none-any.whl (29 kB)
Downloading
aiohttp-3.11.14-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.7
MB)
                         1.7/1.7 MB
87.4 MB/s eta 0:00:00
Downloading certifi-2025.1.31-py3-none-any.whl (166 kB)
Downloading charset_normalizer-3.4.1-cp311-cp311-
```

```
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (143 kB)
Downloading idna-3.10-py3-none-any.whl (70 kB)
Downloading urllib3-2.3.0-py3-none-any.whl (128 kB)
Downloading beautifulsoup4-4.13.3-py3-none-any.whl (186 kB)
Downloading aiohappyeyeballs-2.6.1-py3-none-any.whl (15 kB)
Downloading aiosignal-1.3.2-py2.py3-none-any.whl (7.6 kB)
Downloading attrs-25.3.0-py3-none-any.whl (63 kB)
Downloading frozenlist-1.5.0-cp311-cp311-
manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_6
4.whl (274 kB)
Downloading
multidict-6.2.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (133
kB)
Downloading
propcache-0.3.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (231
kB)
Downloading PySocks-1.7.1-py3-none-any.whl (16 kB)
Downloading soupsieve-2.6-py3-none-any.whl (36 kB)
Downloading
yarl-1.18.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (344 kB)
Building wheels for collected packages: jsbeautifier, stop_words
  Building wheel for jsbeautifier (pyproject.toml) … done
  Created wheel for jsbeautifier: filename=jsbeautifier-1.15.1-py3-none-any.whl
size=94751
sha256=d6069099846fb80b388cb5b892362e4a9d444483290d3b749cf9a406b0b3865f
  Stored in directory: /root/.cache/pip/wheels/32/eb/d8/07a66b0cf535ed05084d0b9a
920b89e35f20512a5e28e615ca
  Building wheel for stop_words (setup.py) … done
  Created wheel for stop_words: filename=stop_words-2018.7.23-py3-none-any.whl
size=32940
sha256=d73f5185f12c98302f7a04d85963471462e3c54fe5a20d95553d7bcdece5f16e
  Stored in directory: /root/.cache/pip/wheels/8f/a5/51/a5405e1da5d178491b79d12c
c81b6cb9bb14fe2c8c632eba70
Successfully built jsbeautifier stop_words
Installing collected packages: stop_words, sortedcontainers, pytz, editorconfig,
urllib3, tzdata, tqdm, threadpoolctl, soupsieve, six, safetensors, regex,
pyyaml, PySocks, propcache, packaging, numpy, multidict, lxml, joblib, idna,
frozenlist, click, charset-normalizer, certifi, attrs, aiohappyeyeballs, yarl,
scipy, sacremoses, requests, python-dateutil, pystempel, lightning-utilities,
jsbeautifier, beautifulsoup4, aiosignal, scikit_learn, pandas, huggingface-hub,
aiohttp, torchmetrics, tokenizers, pyreadr, gdown, transformers,
pytorch_lightning, napkinxc
Successfully installed PySocks-1.7.1 aiohappyeyeballs-2.6.1 aiohttp-3.11.14
aiosignal-1.3.2 attrs-25.3.0 beautifulsoup4-4.13.3 certifi-2025.1.31 charset-
normalizer-3.4.1 click-8.0.3 editorconfig-0.17.0 frozenlist-1.5.0 gdown-5.2.0
huggingface-hub-0.29.3 idna-3.10 joblib-1.4.2 jsbeautifier-1.15.1 lightning-
utilities-0.14.1 lxml-5.3.0 multidict-6.2.0 napkinxc-0.7.1 numpy-1.24.3
packaging-24.2 pandas-2.2.2 propcache-0.3.0 pyreadr-0.5.2 pystempel-1.2.0
```

```
python-dateutil-2.9.0.post0 pytorch_lightning-2.5.0.post0 pytz-2025.1
pyyaml-6.0.2 regex-2024.11.6 requests-2.32.3 sacremoses-0.1.1 safetensors-0.5.3
scikit_learn-1.2.2 scipy-1.15.2 six-1.17.0 sortedcontainers-2.4.0 soupsieve-2.6
stop_words-2018.7.23 threadpoolctl-3.6.0 tokenizers-0.19.1 torchmetrics-1.4.1
tqdm-4.66.5 transformers-4.44.2 tzdata-2025.1 urllib3-2.3.0 yarl-1.18.3
```

[ ]:

```
import sys
sys.path.append("classifier/lib/python3.11/site-packages")
```

Download multilingual models (around 20 min for each file)

- top-down (https://repod.icm.edu.pl/api/access/datafile/50892)
- bottom-up (https://repod.icm.edu.pl/api/access/datafile/50893)

[ ]:

```
!wget https://repod.icm.edu.pl/api/access/datafile/50892 -O top-down.zip
```

```
--2025-03-18 09:38:11--  https://repod.icm.edu.pl/api/access/datafile/50892
Resolving repod.icm.edu.pl (repod.icm.edu.pl)… 213.135.60.199
Connecting to repod.icm.edu.pl (repod.icm.edu.pl)|213.135.60.199|:443…
connected.
HTTP request sent, awaiting response… 200 OK
Length: 2117465081 (2.0G) [application/zip]
Saving to: ‘top-down.zip’

top-down.zip        100%[===================>]   1.97G  4.24MB/s    in 15m 17s

2025-03-18 09:53:30 (2.20 MB/s) - ‘top-down.zip’ saved [2117465081/2117465081]
```

[ ]:

```
!wget https://repod.icm.edu.pl/api/access/datafile/50893  -O bottom-up.zip
```

```
--2025-03-18 09:54:44--  https://repod.icm.edu.pl/api/access/datafile/50893
Resolving repod.icm.edu.pl (repod.icm.edu.pl)… 213.135.60.199
Connecting to repod.icm.edu.pl (repod.icm.edu.pl)|213.135.60.199|:443…
connected.
HTTP request sent, awaiting response… 200 OK
Length: 3049932246 (2.8G) [application/zip]
Saving to: ‘bottom-up.zip’

bottom-up.zip       100%[===================>]   2.84G  3.31MB/s    in 16m 5s

2025-03-18 10:10:50 (3.01 MB/s) - ‘bottom-up.zip’ saved [3049932246/3049932246]
```

[ ]:

```
!unzip -l top-down.zip
```

```
Archive:  top-down.zip
  Length      Date    Time    Name
---------  ---------- -----   ----
        0  2024-10-23 10:26   transformer-multi-top-2024-sample-20/
        0  2024-10-25 06:48   transformer-multi-top-2024-sample-20/ckpts/
3392393018  2024-10-24 02:17   transformer-multi-
top-2024-sample-20/ckpts/epoch=4-val_loss=0.31984.ckpt
   319109  2024-10-23 06:26   transformer-multi-top-2024-sample-20/hierarchy.bin
       88  2024-10-23 06:26   transformer-multi-
top-2024-sample-20/transformer_arch.bin
---------                     -------
3392712215                     5 files
```

[ ]:
```
!unzip top-down.zip
```

```
Archive:  top-down.zip
   creating: transformer-multi-top-2024-sample-20/
   creating: transformer-multi-top-2024-sample-20/ckpts/
  inflating: transformer-multi-
top-2024-sample-20/ckpts/epoch=4-val_loss=0.31984.ckpt
  inflating: transformer-multi-top-2024-sample-20/hierarchy.bin
  inflating: transformer-multi-top-2024-sample-20/transformer_arch.bin
```

[ ]:
```
!unzip bottom-up.zip
```

```
Archive:  bottom-up.zip
   creating: transformer-multi-bottom-2024-sample-20/
  inflating: transformer-multi-bottom-2024-sample-20/hierarchy.bin
  inflating: transformer-multi-bottom-2024-sample-20/transformer_arch.bin
   creating: transformer-multi-bottom-2024-sample-20/ckpts/
  inflating: transformer-multi-
bottom-2024-sample-20/ckpts/epoch=6-val_loss=1.59784.ckpt
  inflating: transformer-multi-bottom-2024-sample-20/config.json
  inflating: transformer-multi-bottom-2024-sample-20/model.safetensors
```

0.3 Loading library and simple prediction

Libraries and functions

[ ]:
```
import numpy as np
import pandas as pd
from job_offers_classifier.load_save import load_to_df, save_obj, load_obj ##␣
 ↪here it is not needed actually
from job_offers_classifier.job_offers_utils import create_hierarchy,␣
 ↪fix_class_str, remove_classes, filter_classes, remap_classes ## here it is␣
 ↪not needed actually
from job_offers_classifier.job_offers_classfier import *
```

Read the models

[ ]:
```
trans_bottom = TransformerJobOffersClassifier(modeling_mode="bottom-up")
trans_bottom.load("transformer-multi-bottom-2024-sample-20")
```

```
trans_top = TransformerJobOffersClassifier(modeling_mode="top-down")
trans_top.load("transformer-multi-top-2024-sample-20")
```

```
Initializing TransformerClassifier with model_name=FacebookAI/xlm-roberta-base,
  output_size=2911, labels_groups=False,
  learning_rate=1e-05, weight_decay=0.01, warmup_steps=50, train_batch_size=8,
eval_batch_size=8, freeze_transformer=False
  loss=flat cross entropy, hidden_dropout=0.0, hierarchy_leaves=2911 …
```

```
/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(
```

```
config.json:   0%|          | 0.00/615 [00:00<?, ?B/s]
```

```
model.safetensors:   0%|          | 0.00/1.12G [00:00<?, ?B/s]
```

```
Initializing TransformerClassifier with model_name=FacebookAI/xlm-roberta-base,
  output_size=3543, labels_groups=True,
  learning_rate=1e-05, weight_decay=0.01, warmup_steps=50, train_batch_size=8,
eval_batch_size=8, freeze_transformer=False
  loss=hierarchical cross entropy, hidden_dropout=0.0, hierarchy_leaves=2911 …
```

Example offers

[ ]:

```
example_offer = [
    "Starszy Statystyk w Ośrodku Metodologii Badań Ludnościowych OSOBA NA TYM␣
↪STANOWISKU: Bierze współudział w rozwoju metodologii estymacji wielkości␣
↪populacji i jej charakterystyk na podstawie zintegrowanych danych Prowadzi␣
↪podstawowe prace projektowo-programistyczne dotyczące estymacji i integracji␣
↪danych Współpracuje z departamentami GUS, urzędami statystycznymi i␣
↪komórkami organizacyjnymi Urzędu w zakresie estymacji wielkości populacji i␣
↪ich charakterystyk na podstawie zintegrowanych danych Uczestniczy we␣
↪współpracy międzynarodowej w dziedzinie estymacji wielkości populacji i ich␣
↪charakterystyk na podstawie zintegrowanych danych Współpracuje przy␣
↪przygotowywaniu danych niezbędnych do realizacji zleceń zewnętrznych␣
↪związanych z udostępnianiem informacji Bierze udział w rozpoznawaniu␣
↪rejestrów administracyjnych jako baz danych ludnościowych Wykształcenie:␣
↪średnie Znajomość języka angielskiego na poziomie komunikatywnym Podstawowa␣
↪znajomość metod statystycznych ze szczególnym uwzględnieniem metody␣
↪reprezentacyjnej Podstawowa znajomość pakietów statystycznych R lub Julia␣
↪(ewentualnie Python) Umiejętność argumentowania Rzetelność Organizacja pracy␣
↪i orientacja na osiąganie celów Wykorzystywanie wiedzy i doskonalenie␣
↪zawodowe Współpraca Komunikacja interpersonalna Wykształcenie: wyższe␣
↪profilowane (matematyka, informatyka, ekonomia) Doświadczenie zawodowe co␣
↪najmniej 1 rok w obszarze metodologii badań statystycznych Znajomość␣
↪programu badań statystycznych statystyki publicznej Myślenie analityczne␣
↪Kreatywność",
    "Specjalista w Ośrodku Metodologii Badań Ludnościowych Koordynuje prace␣
↪związane z projektami realizowanymi w Ośrodku Metodologii Badań␣
↪Ludnościowych Przygotowywuje dokumenty związane z realizacją projektów␣
↪Współpracuje z departamentami GUS Wykształcenie: wyższe Znajomość języka␣
↪angielskiego na poziomie komunikatywnym Umiejętność zarządzania projektami␣
↪Umiejętność argumentowania Rzetelność Organizacja pracy i orientacja na␣
↪osiąganie celów Wykorzystywanie wiedzy i doskonalenie zawodowe Współpraca␣
↪Komunikacja pisemna Komunikacja interpersonalna Wykształcenie: wyższe␣
↪profilowane (ekonomia lub zarządzanie) Doświadczenie zawodowe co najmniej 1␣
↪rok w obszarze metodologii badań statystycznych Znajomość języka␣
↪angielskiego na poziomie bardzo dobrym Znajomość programu badań␣
↪statystycznych statystyki publicznej Myślenie analityczne Kreatywność",
    "We are seeking a highly skilled AI Solutions Engineer to join our team.␣
  ↪The ideal candidate will have extensive experience in machine learning (ML)␣
  ↪and Generative AI, focusing on both the strategic and tactical development␣
  ↪of AI solutions. You will architect and deploy advanced AI systems,␣
  ↪particularly around Generative AI, leveraging your deep technical expertise␣
  ↪to build innovative products, solve complex problems and guide AI-driven␣
  ↪strategies for the business. Bachelor’s or Master’s degree in Computer␣
  ↪Science, Machine Learning, or a related field (PhD preferred). 5+ years of␣
  ↪experience in building and deploying machine learning models, with proven␣
  ↪experience in Generative AI. Expertise in Python and ML frameworks, and deep␣
  ↪learning architectures for generative models. Proven experience in cloud␣
  ↪platforms such as AWS or Azure for AI solution deployment. Experience in␣
  ↪developing products utilizing Generative AI models, such as image␣
  ↪generation, natural language generation, or synthetic data generation.␣
  ↪Strong understanding of ML algorithms (supervised, unsupervised,␣
  ↪reinforcement learning) and Generative AI techniques. Experience in scaling␣
  ↪generative models and deploying them in production environments. Excellent␣
  ↪communication skills to articulate technical concepts to non-technical␣
  ↪stakeholders. Experience in NLP, computer vision, and deploying large␣
  ↪language models (LLMs) for content generation - is an advantage. Hands-on␣
  ↪experience with fine-tuning LLMs - is an advantage. Experience with MLOps␣
  ↪frameworks and version control of models - is nice to have. Knowledge of␣
  ↪product design principles using AI and ability to drive cross-functional AI␣
  ↪projects with a generative component - is an asset.",
    "                                                  /               ,   ␣
  ↪                 .                        /                 ,             ␣
  ↪                                   Python,          ,                   ␣
  ↪         .                                ,             ,                 ␣
  ↪            (Gen)AI.                                  ,             ␣
  ↪    -   .       ,                                      ,                  ␣
  ↪LLM.                   ,                          .                   ␣
  ↪Python,                                             (git,         ).      ␣
  ↪                , GenAI            .                                       ␣
  ↪     AI            .                                                   .     ␣
  ↪          ’       ,               .      5                                .␣
  ↪                    ,       Python.                                .        ␣
  ↪  LLM  RAG                (Langchain, LLamaIndex   ).                     ␣
  ↪NLP (Natural Language Processing).                          .             ␣
  ↪              .                         .                                 ␣
  ↪    "
]
```

Prediction of the 1-digit occupation code

```
[ ]: pred_result_bottom = trans_bottom.predict(
         example_offer, # text data for which to make a prediction
```

```
    output_level=0, # the layer of the hierarchy for which to make a prediction (add a number from 0 - the first level, or 'last' if predicting for the last)
    top_k=None, # how many classes with the highest probability to predict, if None is returned the probability for all classes on the selected level of the hierarchy
    format='dataframe' # defines the format in which the result will be returned, possible options are 'array' - numpy.ndarray or 'dataframe' - pandas.DataFrame
)
```

```
Initializing TransformerDataModule with model_name=FacebookAI/xlm-roberta-base,
max_seq_length=512, train/eval_batch_size=8/8, num_workers=2 …
Setting up TransformerDataModule …

tokenizer_config.json:   0%|          | 0.00/25.0 [00:00<?, ?B/s]

sentencepiece.bpe.model:   0%|          | 0.00/5.07M [00:00<?, ?B/s]

tokenizer.json:   0%|          | 0.00/9.10M [00:00<?, ?B/s]

/content/classifier/lib/python3.11/site-
packages/lightning_fabric/connector.py:572: `precision=16` is supported for
historical reasons but its usage is discouraged. Please set your precision to
16-mixed instead!
INFO:pytorch_lightning.utilities.rank_zero:Using 16bit Automatic Mixed Precision
(AMP)
INFO:pytorch_lightning.utilities.rank_zero:GPU available: True (cuda), used:
True
INFO:pytorch_lightning.utilities.rank_zero:TPU available: False, using: 0 TPU
cores
INFO:pytorch_lightning.utilities.rank_zero:HPU available: False, using: 0 HPUs

Starting predicting with TransformerClassifier …

INFO:pytorch_lightning.utilities.rank_zero:Restoring states from the checkpoint
path at transformer-multi-
bottom-2024-sample-20/ckpts/epoch=6-val_loss=1.59784.ckpt
INFO:pytorch_lightning.utilities.migration.utils:Lightning automatically
upgraded your loaded checkpoint from v1.9.5 to v2.5.0.post0. To apply the
upgrade to your files permanently, run `python -m
pytorch_lightning.utilities.upgrade_checkpoint transformer-multi-
bottom-2024-sample-20/ckpts/epoch=6-val_loss=1.59784.ckpt`
/content/classifier/lib/python3.11/site-
packages/pytorch_lightning/callbacks/model_checkpoint.py:362: The dirpath has
changed from 'models-paper/multi/transformer-multi-bottom-2024-sample-20/ckpts'
to '/content/transformer-multi-
bottom-2024-sample-20/ckpts/epoch=6-val_loss=1.59784.ckpt', therefore
`best_model_score`, `kth_best_model_path`, `kth_value`, `last_model_path` and
`best_k_models` won't be reloaded. Only `best_model_path` will be reloaded.
INFO:pytorch_lightning.accelerators.cuda:LOCAL_RANK: 0 - CUDA_VISIBLE_DEVICES:
```

```
[0]
INFO:pytorch_lightning.utilities.rank_zero:Loaded model weights from the
checkpoint at transformer-multi-
bottom-2024-sample-20/ckpts/epoch=6-val_loss=1.59784.ckpt
```

```
Predicting: |          | 0/? [00:00<?, ?it/s]
```

[ ]:

```
pred_result_top = trans_top.predict(
    example_offer,
    output_level=0,
    top_k=None,
    format="dataframe"
)
```

```
Initializing TransformerDataModule with model_name=FacebookAI/xlm-roberta-base,
max_seq_length=512, train/eval_batch_size=8/8, num_workers=2 …
Setting up TransformerDataModule …
```

```
INFO:pytorch_lightning.utilities.rank_zero:Using 16bit Automatic Mixed Precision
(AMP)
INFO:pytorch_lightning.utilities.rank_zero:GPU available: True (cuda), used:
True
INFO:pytorch_lightning.utilities.rank_zero:TPU available: False, using: 0 TPU
cores
INFO:pytorch_lightning.utilities.rank_zero:HPU available: False, using: 0 HPUs
INFO:pytorch_lightning.utilities.rank_zero:Restoring states from the checkpoint
path at transformer-multi-top-2024-sample-20/ckpts/epoch=4-val_loss=0.31984.ckpt
```

```
Starting predicting with TransformerClassifier …
```

```
INFO:pytorch_lightning.utilities.migration.utils:Lightning automatically
upgraded your loaded checkpoint from v1.9.5 to v2.5.0.post0. To apply the
upgrade to your files permanently, run `python -m
pytorch_lightning.utilities.upgrade_checkpoint transformer-multi-
top-2024-sample-20/ckpts/epoch=4-val_loss=0.31984.ckpt`
/content/classifier/lib/python3.11/site-
packages/pytorch_lightning/callbacks/model_checkpoint.py:362: The dirpath has
changed from 'models-paper/multi/transformer-multi-top-2024-sample-20/ckpts' to
'/content/transformer-multi-
top-2024-sample-20/ckpts/epoch=4-val_loss=0.31984.ckpt', therefore
`best_model_score`, `kth_best_model_path`, `kth_value`, `last_model_path` and
`best_k_models` won't be reloaded. Only `best_model_path` will be reloaded.
INFO:pytorch_lightning.accelerators.cuda:LOCAL_RANK: 0 - CUDA_VISIBLE_DEVICES:
[0]
INFO:pytorch_lightning.utilities.rank_zero:Loaded model weights from the
checkpoint at transformer-multi-
top-2024-sample-20/ckpts/epoch=4-val_loss=0.31984.ckpt
```

```
Predicting: |          | 0/? [00:00<?, ?it/s]
```

Predictions for each 1-digit occupation code for both models (2 two rows: bottop-up model, next two: top-down model).

[ ]:

```
pd.concat([pred_result_bottom,pred_result_top])
```

[ ]:

```
          0         1         2         3         4         5         6  \
0  0.000042  0.006040  0.404478  0.237416  0.313389  0.007069  0.001013
1  0.000012  0.063933  0.609985  0.038708  0.259444  0.008965  0.000519
2  0.000014  0.004593  0.960370  0.011890  0.002996  0.003023  0.000908
3  0.000012  0.002445  0.970239  0.012542  0.005935  0.001242  0.000462
0  0.000312  0.002083  0.828137  0.052120  0.114958  0.000511  0.000613
1  0.000269  0.123873  0.634011  0.002438  0.236913  0.001435  0.000465
2  0.000951  0.010495  0.983327  0.000168  0.000621  0.001219  0.001900
3  0.000598  0.001582  0.992644  0.000637  0.003383  0.000294  0.000585

          7         8         9
0  0.018153  0.009070  0.003331
1  0.010591  0.005048  0.002794
2  0.008809  0.005000  0.002397
3  0.003562  0.002545  0.001016
0  0.000448  0.000461  0.000356
1  0.000127  0.000087  0.000384
2  0.000178  0.000513  0.000628
3  0.000009  0.000090  0.000178
```

Prediction of 6-digits occupation code (top 3)

[ ]:

```
pred_top3_bottom = trans_bottom.predict(
    example_offer,
    output_level="last",
    top_k=3,
    format="dataframe"
)

pred_top3_top = trans_top.predict(
    example_offer,
    output_level="last",
    top_k=3,
    format="dataframe"
)
```

```
Initializing TransformerDataModule with model_name=FacebookAI/xlm-roberta-base,
max_seq_length=512, train/eval_batch_size=8/8, num_workers=2 …
Setting up TransformerDataModule …
```

```
/content/classifier/lib/python3.11/site-
packages/lightning_fabric/connector.py:572: `precision=16` is supported for
historical reasons but its usage is discouraged. Please set your precision to
16-mixed instead!
INFO:pytorch_lightning.utilities.rank_zero:Using 16bit Automatic Mixed Precision
```

```
(AMP)
INFO:pytorch_lightning.utilities.rank_zero:GPU available: True (cuda), used:
True
INFO:pytorch_lightning.utilities.rank_zero:TPU available: False, using: 0 TPU
cores
INFO:pytorch_lightning.utilities.rank_zero:HPU available: False, using: 0 HPUs
INFO:pytorch_lightning.utilities.rank_zero:Restoring states from the checkpoint
path at transformer-multi-
bottom-2024-sample-20/ckpts/epoch=6-val_loss=1.59784.ckpt

Starting predicting with TransformerClassifier …

INFO:pytorch_lightning.utilities.migration.utils:Lightning automatically
upgraded your loaded checkpoint from v1.9.5 to v2.5.0.post0. To apply the
upgrade to your files permanently, run `python -m
pytorch_lightning.utilities.upgrade_checkpoint transformer-multi-
bottom-2024-sample-20/ckpts/epoch=6-val_loss=1.59784.ckpt`
/content/classifier/lib/python3.11/site-
packages/pytorch_lightning/callbacks/model_checkpoint.py:362: The dirpath has
changed from 'models-paper/multi/transformer-multi-bottom-2024-sample-20/ckpts'
to '/content/transformer-multi-
bottom-2024-sample-20/ckpts/epoch=6-val_loss=1.59784.ckpt', therefore
`best_model_score`, `kth_best_model_path`, `kth_value`, `last_model_path` and
`best_k_models` won't be reloaded. Only `best_model_path` will be reloaded.
INFO:pytorch_lightning.accelerators.cuda:LOCAL_RANK: 0 - CUDA_VISIBLE_DEVICES:
[0]
INFO:pytorch_lightning.utilities.rank_zero:Loaded model weights from the
checkpoint at transformer-multi-
bottom-2024-sample-20/ckpts/epoch=6-val_loss=1.59784.ckpt

Predicting: |          | 0/? [00:00<?, ?it/s]

Initializing TransformerDataModule with model_name=FacebookAI/xlm-roberta-base,
max_seq_length=512, train/eval_batch_size=8/8, num_workers=2 …
Setting up TransformerDataModule …

INFO:pytorch_lightning.utilities.rank_zero:Using 16bit Automatic Mixed Precision
(AMP)
INFO:pytorch_lightning.utilities.rank_zero:GPU available: True (cuda), used:
True
INFO:pytorch_lightning.utilities.rank_zero:TPU available: False, using: 0 TPU
cores
INFO:pytorch_lightning.utilities.rank_zero:HPU available: False, using: 0 HPUs
INFO:pytorch_lightning.utilities.rank_zero:Restoring states from the checkpoint
path at transformer-multi-top-2024-sample-20/ckpts/epoch=4-val_loss=0.31984.ckpt

Starting predicting with TransformerClassifier …

INFO:pytorch_lightning.utilities.migration.utils:Lightning automatically
upgraded your loaded checkpoint from v1.9.5 to v2.5.0.post0. To apply the
upgrade to your files permanently, run `python -m
```

```
pytorch_lightning.utilities.upgrade_checkpoint transformer-multi-
top-2024-sample-20/ckpts/epoch=4-val_loss=0.31984.ckpt`
/content/classifier/lib/python3.11/site-
packages/pytorch_lightning/callbacks/model_checkpoint.py:362: The dirpath has
changed from 'models-paper/multi/transformer-multi-top-2024-sample-20/ckpts' to
'/content/transformer-multi-
top-2024-sample-20/ckpts/epoch=4-val_loss=0.31984.ckpt', therefore
`best_model_score`, `kth_best_model_path`, `kth_value`, `last_model_path` and
`best_k_models` won't be reloaded. Only `best_model_path` will be reloaded.
INFO:pytorch_lightning.accelerators.cuda:LOCAL_RANK: 0 - CUDA_VISIBLE_DEVICES:
[0]
INFO:pytorch_lightning.utilities.rank_zero:Loaded model weights from the
checkpoint at transformer-multi-
top-2024-sample-20/ckpts/epoch=4-val_loss=0.31984.ckpt
```

```
Predicting: |          | 0/? [00:00<?, ?it/s]
```

Results for top 3 along with the probabilities

```
[ ]: pd.concat([pred_top3_bottom,pred_top3_top])
```

```
[ ]:   class_1 class_2 class_3    prob_1    prob_2    prob_3
     0  431201  331401  252102  0.288475  0.193673  0.170916
     1  242290  441990  411090  0.114219  0.080363  0.079738
     2  251908  252102  251201  0.768940  0.035406  0.031738
     3  251908  252102  252990  0.574634  0.329984  0.011291
     0  212004  216590  431201  0.404771  0.082207  0.058343
     1  242112  411090  242102  0.273998  0.130173  0.077134
     2  251908  251201  251202  0.386948  0.224004  0.161463
     3  252102  251908  252990  0.492204  0.336939  0.033593
```