

## Documentation

This code seeks to apply a version of BERT called "ClinicalBERT" to data that originates from Indiana chest X-ray dataset. ClinicalBERT was initially developed by Kexin Huang, Jaan Altosaar, Rajesh Ranganath as outlined in their article "[ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission](#)", and was developed by training BERT on MIMIC III hospital readmission data. The code in this project follows that described in an article by Nwamake Imasogie, where she applies a similar analysis to 2-day and 3-day clinical notes, instead of the discharge notes used in the original article.

## Code

The code for this repo is contained in two files.

prep\_caml\_data.ipynb → for preprocessing the code and conducting classical NLP ML analysis

CAMLBERT\_final.ipynb → for running ClinicalBERT on the data

There are flags at the top of the code for switching between Mac and Colab platforms, though Mac settings have not been tested in a while and may not work.

The original code provided by Nwamake Imasogie can be found at [this repo](#) (obtained after contacting her ... the links in her original articles are broken). The NLP analysis is different from that conducted by the author, and can be considered to be **"additional modeling" (= ablation study)**.

## BERT Model and Code

The pretrained ClinicalBERT Model parameters are provided under the Model folder, but the actual BERT bin file is too large to store here and needs to be uploaded from [https://drive.google.com/drive/folders/1X\\_oOiKWE5WRebNDAYniafuyycQZYkQYXu](https://drive.google.com/drive/folders/1X_oOiKWE5WRebNDAYniafuyycQZYkQYXu). The file structure needs to be properly maintained, as suggested by the file paths at the top of the Jupyter Notebook CAMLBERT\_final.ipynb.

| File                                  | Location                  |
|---------------------------------------|---------------------------|
| file_utils.py                         | top level directory       |
| modeling_readmission.py               | top level directory       |
| model (folder) / readmission (folder) | top level directory       |
| pytorch_model.bin                     | in the readmission folder |
| bert_config.json                      | in the readmission folder |

## Data Location

The chest X-ray data provided for this project is contained in this repo (clean\_caml.csv). The original train / test splits are also provided for comparison (caml\_train.csv and caml\_test.csv).

## Original Papers

This project seeks to apply the general approach taken by Nwamake Imasogie in her paper [ClinicalBERT: Using a Deep Learning Transformer Model to Predict Hospital Readmission](#). The code repo is available [here](#) and a related article on how she preprocessed the data and applied classical NLP techniques and ML analysis can be found at [Predicting Hospital Readmission Using NLP](#). The link to the code base for that article is broken, and the author did not have time to send it to me (she is on maternity leave), but promised to send it at a future date.

## Dependencies

## Data Download Instructions

No need to download. Everything is contained in this repo.