# 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 most important code for this repo is contained in two files.
prep_caml_data.ipynb → for preprocessing the data and conducting classical NLP ML analysis (logistic regression with bag of words and TF-IDF)
CAMLBERT_final.ipynb → for running ClinicalBERT on the data

There are flags at the top of the CAMLBERT_final.ipynb for switching between Mac and Colab platforms, though Mac settings have not been tested in a while and may not work. prep_caml_data.ipynb was run on a Mac and would have to be adapted for Colab.

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).

## CAML 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). The “binary polarity reversed” version of clean_caml.csv is rev_clean_caml.csv, and is kept in the top level directory.

## 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

Python 3.7.13 (higher versions also probably work)
transformers 4.18.0
pytorch 1.11.0+cu113
boto3 1.™22.9
Note: the last three dependencies are already installed within the Jupyter notebook. The Python version seems to have come by default with the notebook.

### Data Download Instructions: BERT Model and Code

Almost everything you need to run this repo is here in the repo itself. For instance, the pretrained ClinicalBERT Model parameters (bert_config.json) are provided under the model folder. However, the actual BERT bin file (pytorch_model.bin) is simply too large to store in this repo. It is currently available from this Google Drive (courtesy of Nwamake Imasogie). https://drive.google.com/drive/folders/1X_oOiKWE5WRebNDAyniafuycQZYkQYXu.

### Note on File Locations

File locations and requirements are described below. However, your file structure may differ. Look at the cells on the top of CAMLBERT_final.ipynb for guidance on how to locate the files on your system, and adjust accordingly.

### Data File Locations and Purposes

| Data | Location | Needed by |
|---|---|---|
| clean_caml.csv | top level directory | CAMLBERT_final.ipynb<br>reverse_caml_label.ipynb |
| rev_clean_caml.csv | top level directory | camlbert_final.ipynb |
| caml_test.csv | top level directory | prep_caml_data.ipynb |
| caml_train.csv | top level directory | prep_caml_data.pynb |

### Notebooks

| Notebook | Purpose |
|---|---|
| CAMLBERT_final.ipynb | the main one for running CAMLBERT |
| prep_caml_data.ipynb | cleans the caml data in prep for camlbert |
| rev_caml_label.ipynb | reverses the polarity of the label |
| dlhc_project_bert.ipynb | notebook for preparing and modeling data assuming you have MIMIC-III; this was submitted for draft, because at the time I thought I would be gaining access to MIMIC-III data |

### File Locations

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