# DSCI 691 - NLP with Deep Learning, Spring 2024

Contributors:

\* Samson Adeniyi (saa384@drexel.edu)

\* Alec Peterson (ap3842@drexel.edu)

\* Morgan Purcell (mrp366@drexel.edu)

# Abstract

# Introduction

This repo contains files to develop sentence-transformer models for producing semantic rich embeddings for clinical texts, exploring different control and experimental conditions to examine the effects of fine-tuning transformer models on general natural language inference (NLI) and medical NLI sentence pairs.

## Files:

\* DSCI\_691\_Model\_Finetuning.ipynb

\* DSCI\_691\_Mean\_Reciprocal\_Rank.ipynb

\* DSCI\_691\_MedNLI\_Classification.ipynb

\* DSCI\_691\_BIOSESS\_Comparison.ipynb

# Background

## Models

Clinical-BERT (https://huggingface.co/medicalai/ClinicalBERT) is a fine-tuned version of the popular BERT encoder transformer architecture, trained on MIMIC-III (Johnson et al., 2016), a large corpus of de-identified clinical records. Clinical-BERT, specifically a version that used the knowledge-distilled DistilBERT model, was used as a base model for our work. For control comparisons to examine the effect of fine-tuning alone, the distilbert-base-uncased model was used.

## Datasets

Two datasets were used for fine-tuning:

\* Stanford NLI (SNLI, https://nlp.stanford.edu/projects/snli/) + Multiple-Genre NLI (MultiNLI, https://cims.nyu.edu/~sbowman/multinli/) combined, also referred to as “AllNLI” - ~1,000,000 samples.

\* MedNLI (https://jgc128.github.io/mednli/) - ~15,000 samples. This dataset consisted of a development, training, and test split. The training split (~11,000 samples) was used for fine-tuning, while the test split (~1400 samples) was used for evaluation

These datasets consist of sets of sentence pairs consisting of a “premise” sentence, and a “hypothesis” sentence. Each pair has an associated label of “entailment”, “neutral”, or “contradiction” referring to how semantically similar the sentences are.

Another dataset, BIOSESS (https://huggingface.co/datasets/tabilab/biosses) was used for evaluation. This dataset consisted of 100 biomedical sentence pairs and a score for each pair ranging from 0 (no relation) to 4 (equivalent) indicating the similarity of the sentences, as assessed by five human annotators.

# Methods

## Hypothesis

Because labeled clinical textual datasets are resource-intensive and challenging to obtain compared to general texts, we were interested in examining the embedding quality of a transformer model with clinical knowledge and pre-training, such as Clinical-BERT, when fine-tuned on \*general\* NLI data compared to \*clinical-specific\* NLI data.

## Conditions

To this end, several control and experimental conditions were devised to assess the impact of NLI fine-tuning:

\* Control 0 (CTRL0): Negative control, the base ClinicalBERT model with no fine-tuning. Indicates embedding quality from clinical pre-trained knowledge alone.

\* Positive Controls: BERT models (i.e. no clinical pre-trained knowledge) to examine the effect of fine-tuning along

\* Control 1 (CTRL1) - SNLI+MultiNLI

\* Control 2 (CTRL2) - MedNLI

\* Control 3 (CTRL3) - SNLI+MultiNLI followed by MedNLI

\* Experimental Conditions: Clinical-BERT models to examine the impact of fine-tuning a model with clinical pre-trained knowledge

\* Experimental 1 (EXP1) - SNLI+MultiNLI

\* Experimental 2 (EXP2) - MedNLI

\* Experimental 3 (EXP3) - SNLI+MultiNLI followed by MedNLI

## Training

### Sentence Transformer Modules

The `sentence-transformers` library was leveraged to add the required modules to the Hugging Face transformer models previously mentioned. These modules consisted of:

\* Mean Pooling Layer - a layer which averages the token embeddings produced from the base transformer model

\* Multiple Negative Ranking (MNR) Loss - a loss function that, within a batch of sample pair embeddings for a given sentence a\_i and other sentences b\_j:

1) calculates the similarity between all combinations of a\_i and b\_j

2) calculates the cross-entropy loss between these similarities and assigned labels, where the semantically similar positive pair a\_i and b\_i within a given batch are assigned the "ideal" label

3) Backpropagation proceeds to optimize cross-entropy loss and thus make the semantically similar a\_i and b\_i close together in the embedding space, and all other b\_j sentences farther away

To align with the intent of MNR loss and standard practice, in effect only the "entailment" and "contradiction" samples were used.

### Training Hyperparameters

Training was performed in DSCI\_691\_Model\_Finetuning.ipynb and used the following hyperparameters:

\* `batch\_size` = 16, mostly due to GPU memory limitations

\* `warmup\_steps` = 10% of training data (a good rule-of-thumb from sentence-transformers authors at sbert.net)

\* `num\_epochs`:

\* = 1 for the SNLI+MultiNLI dataset, primarily due to GPU resource limitations

\* = 10 for the MedNLI training dataset. Since the MedNLI train set was ~1/10 the size of the SNLI+MultiNLI dataset, 10x the epochs were used so that the model was exposed to the same number of iterations/samples

## Evaluation

1) Mean Reciprocal Rank (MRR)

MRR is a common information retrieval measure. While the MedNLI dataset is not of the ideal format for this assessment, the "entailment" positive pairs from the MedNLI validation split were used. Cosine similarity is calculated between premise sentence a\_i and all other entailment hypothesis sentences b\_j. Assuming uniqueness between all other pairs except a\_i and b\_i, the cosine similarity between a\_i and b\_i should be of rank 1 (or at least close to it). This evaluation. is performed in DSCI\_691\_MedNLI\_Classification.ipynb

2) Classification on MedNLI Test Set

One way of assessing the quality of the produced embeddings is to use them as features for a multi-class classifier. A multi-class classifier neural network was trained on the MedNLI train set by concatenating the embeddings of each premise and hypothesis sentence within a given sentence pair. This evaluation, as well as specifics for the classifier neural network, are shown in DSCI\_691\_MedNLI\_Classification.ipynb

3) BIOSESS Correlation to Gold Labels

A similarity score is calculated between each of BIOSESS sentences. After normalization to 0 - 4 like the target labels, a Pearson correlation is calculated. A correlation coefficient close to 1 would indicate agreement with human assessment.

# Results

1) Mean Reciprocal Rank (MRR)

| Model | MRR |

|-------|---------|

| CTRL0 | 0.015862|

| CTRL1 | 0.012445|

| EXP1 | 0.017658|

| CTRL2 | 0.047981|

| EXP2 | 0.058847|

| CTRL3 | 0.018088|

| EXP3 | 0.064095|

2) Classification on MedNLI Test Set

| Model | Accuracy | Precision | Recall | F1-score |

|-------|----------|-----------|--------|----------|

| CTRL0 | 0.682841 | 0.687191 | 0.682841 | 0.684240 |

| CTRL1 | 0.668073 | 0.673463 | 0.668073 | 0.669328 |

| EXP1 | 0.718003 | 0.719465 | 0.718003 | 0.718617 |

| CTRL2 | 0.661041 | 0.662777 | 0.661041 | 0.658784 |

| EXP2 | 0.720816 | 0.738972 | 0.720816 | 0.715694 |

| CTRL3 | 0.667370 | 0.674723 | 0.667370 | 0.668987 |

| EXP3 | 0.729255 | 0.741342 | 0.729255 | 0.724303 |

![Confusion\_Matrices](./images/confusion\_matrices.png)

3) BIOSESS Correlation to Gold Labels

| Model | Pearson Correlation | P-value |

|-------|---------------------|------------------|

| CTRL0 | 0.574784 | 4.008899e-10 |

| CTRL1 | 0.498372 | 1.316569e-07 |

| EXP1 | 0.669632 | 2.591172e-14 |

| CTRL2 | 0.586494 | 1.439203e-10 |

| EXP2 | 0.517410 | 3.544764e-08 |

| CTRL3 | 0.513374 | 4.713201e-08 |

| EXP3 | 0.590170 | 1.034596e-10 |

# Discussion

Our study revealed that Clinical-BERT models fine-tuned on clinical NLI data (MedNLI) outperformed those fine-tuned solely on general NLI data. The Mean Reciprocal Rank (MRR) results indicated that positive pairs (entailment) generally had higher ranks, though overall MRR values were low, suggesting room for improvement in distinguishing entailment pairs. The low MRR results could be partially attributed to the lack of richness or detail in the “hypothesis” second sentence relative to the “premise” first sentence in a pair. The hypothesis was often a terse summary compared to the description in the premise.

Combining general NLI fine-tuning with subsequent clinical NLI fine-tuning yielded the best performance, highlighting the complementary benefits of both datasets. The BIOSESS evaluation showed strong alignment between our model-generated embeddings and human assessments, as indicated by high Pearson correlation coefficients.

# Conclusion

Fine-tuning transformer models on domain-specific datasets significantly enhances their performance in specialized fields like clinical text processing. Clinical-BERT, when fine-tuned with clinical NLI data, produced superior embeddings and classification results compared to models trained on general NLI data alone. Future work should explore additional domain-specific datasets and fine-tuning techniques to further improve model performance.

# References

1. Reimers & Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. <https://doi.org/10.48550/arXiv.1908.10084>. (2019)
2. Devlin et al., BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. <https://arxiv.org/abs/1810.04805>. (2019)
3. Huang et al., ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. <https://doi.org/10.48550/arXiv.1904.05342> (2019)
4. Sanh et al., DistilBERT, a distilled version of BERT: smaller, faster, cheaper, and lighter. <https://arxiv.org/abs/1910.01108>. (2019)
5. Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. <https://doi.org/10.13026/C2XW26>.