Automating the Detection of PHI in Clinical Notes With BERT

Rebecca Golm, rebecca.golm@rutgers.edu, Jorge Ortiz, PhD
Electrical and Computer Engineering Department, Rutgers University, Piscataway, 08854

Abstract

Goal: Improve detection of PHI in clinical notes using BERT

- Modified BERT for NER of PHI
- Pretrained BERT with MIMIC-III database
- Studied the effect of the number of pretraining steps on the model performance
- Evaluated the models using the confusion matrix

Impact: Automation of detection of PHI will decrease the cost of deidentifying medical text increasing its availability to researchers looking to improve the health industry

Introduction

- Natural Language Processing (NLP): applying machine learning to the study of language.
- Named Entity Recognition (NER): identifying important information in text and categorizing it into groups.
- Protected Health Information (PHI): data protected under health data law.

Tools

- BERT(Bidirectional Encoder Representation from Transformers): State of the art model published in October 2018 by Google which performed significantly better on 11 NLP Tasks
  - Uncased BERT Base Model
- i2b2: clinical text with labeled PHI
- MIMIC-III: unlabeled clinical database of Intensive Care Unit Patients

Background

Confusion Matrix:

For binary classification problem:

Actual values

Predicted values	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

$$\text{F1 Score} = \frac{2*\text{precision}*\text{recall}}{\text{precision} + \text{recall}} \qquad \text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

For multiclass classification problem:

- Columns and rows become the different classes
- Precision, recall, and F1 score is per class and averaged
- Accuracy: correctly classified (diagonal) /total number of classifications (sum of all numbers in table)

Methods

Pretraining Method:

Data Preprocessing:

- MIMIC-III Note events file
- Removed Notes where isError is True
- Pasted notes into a text file
- Split file into 13 parts (~300 MB each)
- Run create_pretraining_data.py to prepare the text files for pretraining with BERT

Hyperparameters:

- Number of training steps: 800k
- Save Checkpoint every 20k steps
- Uncased Bert files for checkpoint, config, and vocab
- Otherwise default parameters

Fine Tuning Method 1 (Xavier):

Hyperparameters:

- Output layer weights are Initialized with Xavier Initialization
- Default parameters otherwise

Fine Tuning Method 2 (Default):

Hyperparameters:

- Output layer weights initialized by default
- Default parameters otherwise

Hardware Used: *AWS EC2 Instances*

- P3.2xlarge instance (Tesla V100 GPU)

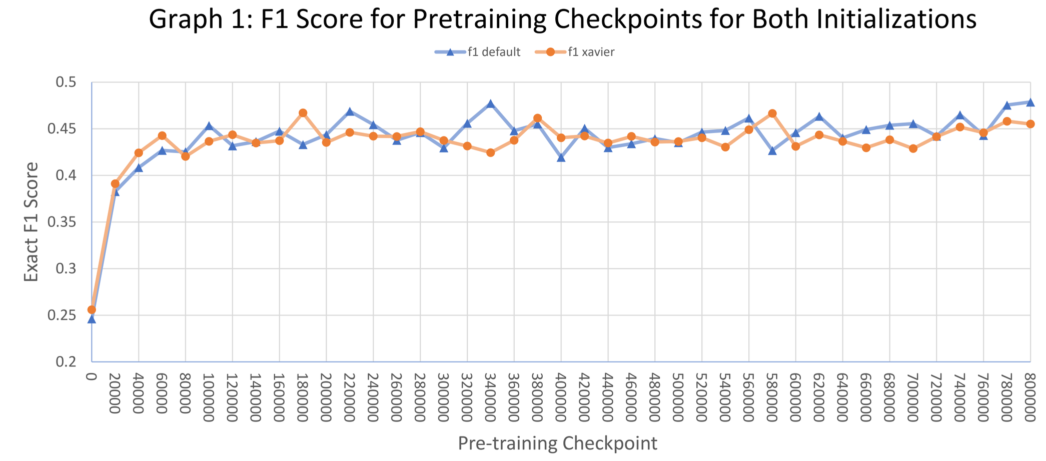
Evaluation:

- Confusion Matrix: F1 Score, Accuracy, Recall, Precision

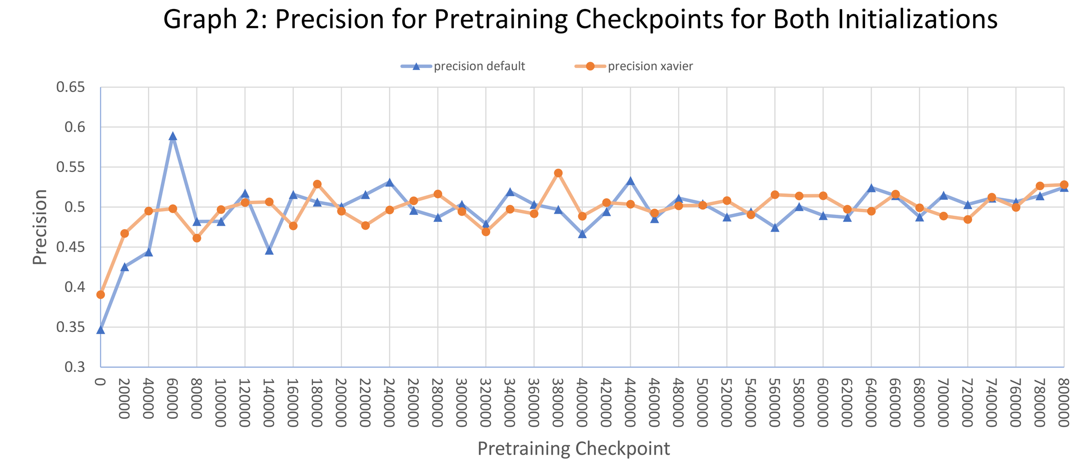
Results

Effect of Output Layer Weight Initialization and Number of Pretraining Steps

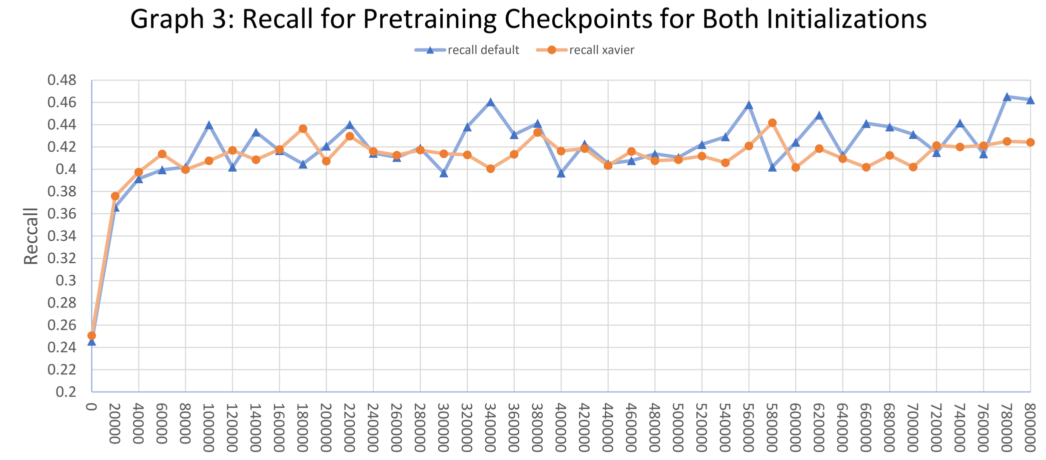
Graph 1: F1 Score for Pretraining Checkpoints for Both Initializations



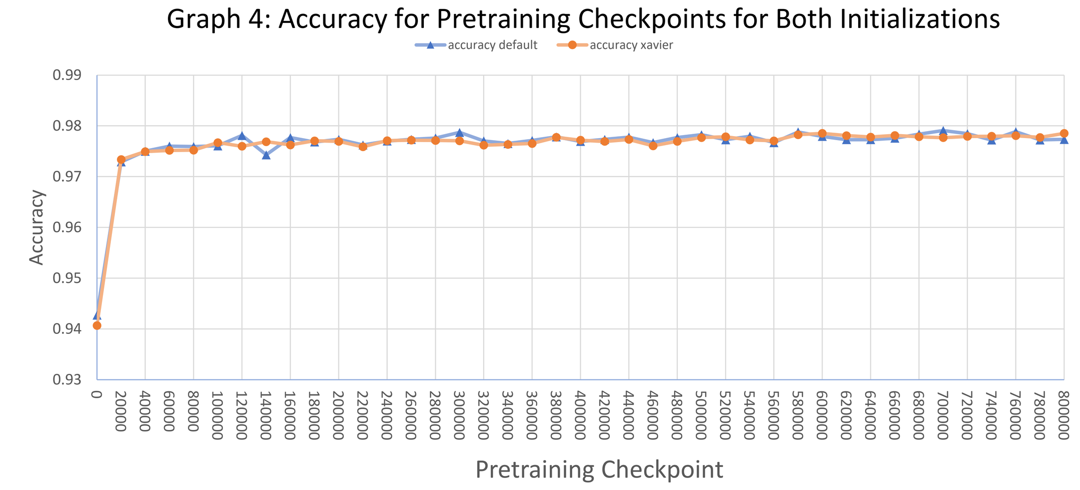
Graph 2: Precision for Pretraining Checkpoints for Both Initializations



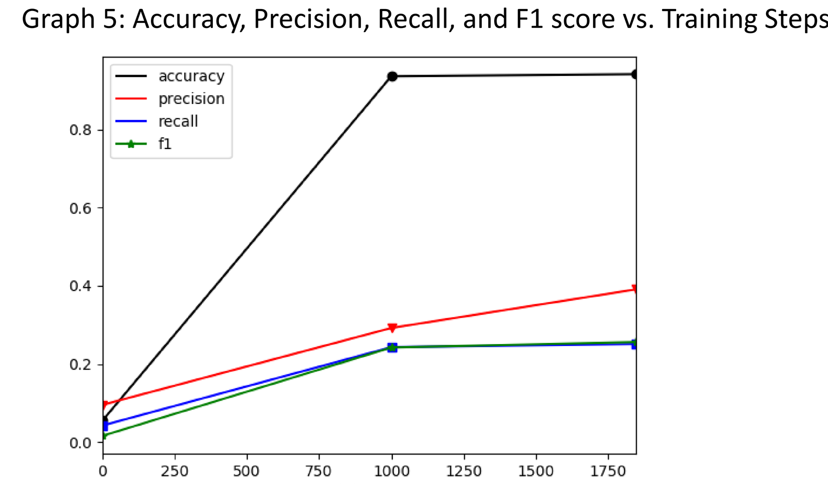
Graph 3: Recall for Pretraining Checkpoints for Both Initializations



Graph 4: Accuracy for Pretraining Checkpoints for Both Initializations



Graph 5: Accuracy, Precision, Recall, and F1 score vs. Training Steps



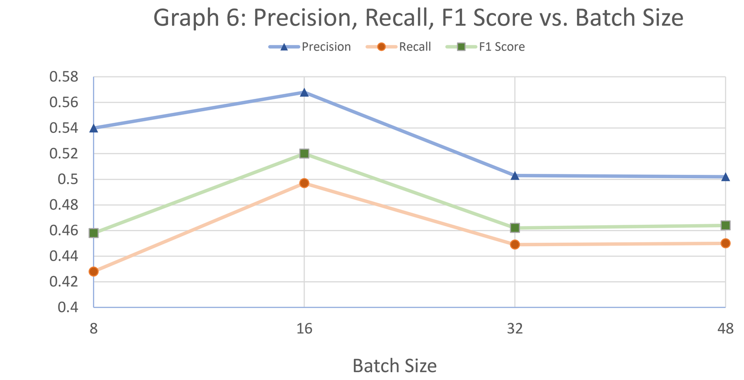
Above: F1 Score, Recall, Precision, and Accuracy versus number of pretraining steps on MIMIC Data for both fine tuning methods

Left: Accuracy, Precision, Recall, and F1 Score versus fine-tuning steps on the published BERT checkpoint for 3 Training Epochs

Results

Effect of Hyperparameters in Finetuning

Graph 6: Precision, Recall, F1 Score vs. Batch Size



Graph 7: Accuracy vs Batch Size

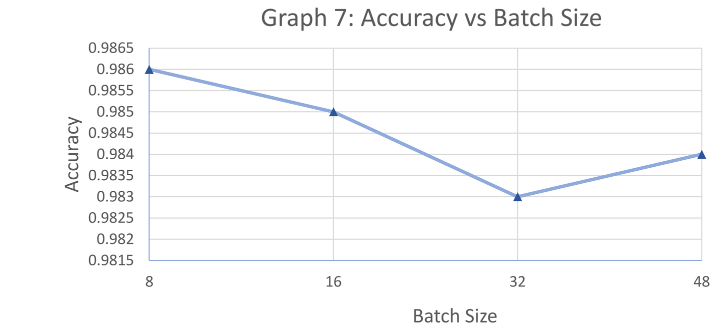


Table 1: Corresponding Epochs for Batch Size Used

Batch Size	8	16	32	48
Training Epochs	3	6	3	5

Above: The corresponding number of epochs for the batch sizes used as hyperparameters in the trainings on the right

Left: Fine tuning evaluation for the above combinations of batch size and epochs without MIMIC pretraining

Discussion

- The Output Layer Initialization has a small impact on performance
  - The default initialization performs on average < 1% better in recal,l F1 Score, and accuracy
  - Xavier initialization performs on average < 0.5% better than the default
- Graph 5 shows an increase in Accuracy, Recall, Precision, and F1 Score with respect to the number of fine tuning training steps
- Graph 6 shows better Precision, Recall, and F1 Score when using Batch Size 16 and 6 Training Epochs compared to other tested hyperparameter combination; however this does not result in the best accuracy
- Graph 7 shows a less than .5% difference in the average of all tested hyperparameter combinations

Future Directions

Improve
- Improve BERT's detection of PHI in Clinical Text (Adjust Hyperparameters)
- Learn from other published BERT models with F1 score > 80%

Apply
- Use the model to identify and remove PHI in clinical text

Share
- Share deidentified medical texts to use for research purposes

Acknowledgements

I want to thank Dr. Jorge Ortiz for his guidance in performing my research this summer. Furthermore, thank you to the Douglass Residential College for enabling my research with the Project SUPER Summer Stipend. Specifically, thank you to Nicole Wodzinski and Kayla Fowler for providing advice and answering all my questions.

References

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Retrieved from https://arxiv.org/abs/1810.04805
- Uzuner Ö., Juo Y, Szolovits P. "Evaluating the state-of-the-art in automatic de-identification". *J Am Med Inform Assoc.* 2007, 14(5):550-63. http://www.jamia.org/cgi/content/abstract/14/5/550.
- MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635
- Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215.full]; 2000 (June 13).

RUTGERS
School of Engineering

RUTGERS
Douglass Residential College