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 FER 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

Background

Confusion Matrix:

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td>Negative</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

F1 Score = 2*precision*recall / precision + recall
Precision = TP / (TP + FP)
Accuracy = (TP + TN) / (TP + TN + FP + FN)

Methods

Pretraining Method:
- 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 law data.

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

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

Discussion

- The Output Layer Initialization has a small impact on performance
  - The default initialization performs on average < 1% better in recall, 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 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 BERT’s detection of PHI in Clinical Text (Adjust Hyperparameters)
- Learn from other published BERT models with F1 score > 80%
- Use the model to identify and remove PHI in clinical text
- 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 Medrano and Kayla Fowler for providing advice and answering all my questions.

References


i2b2: Clinical text with labeled PHI. (2007). i2b2: Clinical text with labeled PHI. Retrieved from http://www.jamia.org/cgi/content/abstract/14/5/550

e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215]

Acknowledgements

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

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)