Comparing Long Sequence Transformer Models for Summarization and Multi-label Classification of Clinical Notes

William Flinchbaugh[1,2], Bishal Neupane[3], Aditya Vadrevu[1], Prashanth Vangari[1], Thasina Tabashum[1], Mark V. Albert[1]

[1]Department of Computer Science and Engineering, University of North Texas, [2]Department of Computer Science, Southern Methodist University,
[3]Department of Computer Science and Engineering, University of Texas at Dallas

Abstract

As more language models become available, we aim to compare how these models can be utilized in the medical field. We specifically focus on transformer models which can take a longer input sequence due to the long nature of clinical notes. For summarization, we compare PEGASUS-X, BigBird-Pegasus, and LongT5. For multi-label classification, we use long sequence transformers to generate a vector representation of the clinical notes and feed that representation into a classification layer. We specifically compare BigBird-RoBERTa and Longformer. We found that the pre-trained BigBird-Pegasus model performed best on the summarization task, but none of the models performed particularly well and need to be fine-tuned. For classification, we found that BigBird-RoBERTa performed significantly better than Longformer by every metric. In the future we plan to fine-tune each of the summarization models. We also plan to compare the classification approaches to traditional machine learning methods.

Summarization

Models:
1. PEGASUS-X
2. BigBird-Pegasus
3. LongT5

Metrics:
1. ROUGE-1 (1-grams)
2. ROUGE-2 (bigrams)
3. ROUGE-L (longest common subsequence)

Classification

Models:
1. BigBird-RoBERTa
2. Longformer

Metrics:
1. Precision
2. Recall
3. F1
4. ROC AUC

Methods

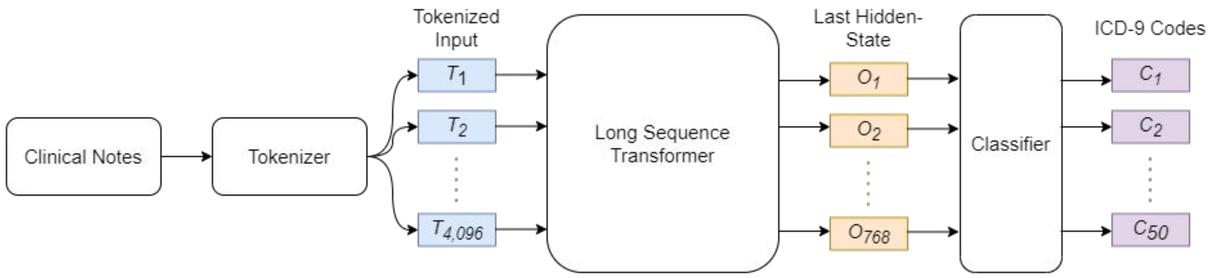

Figure 1. Architecture of the long sequence transformer classification approach.

ICD-9 Code	Health Condition
401.9	Hypertension
38.93	Venous Catheterization
428.0	Congestive Heart Failure
427.31	Atrial Fibrillation
414.01	Coronary Atherosclerosis

Table 1. Top 5 ICD-9 codes and the corresponding health conditions.

Results

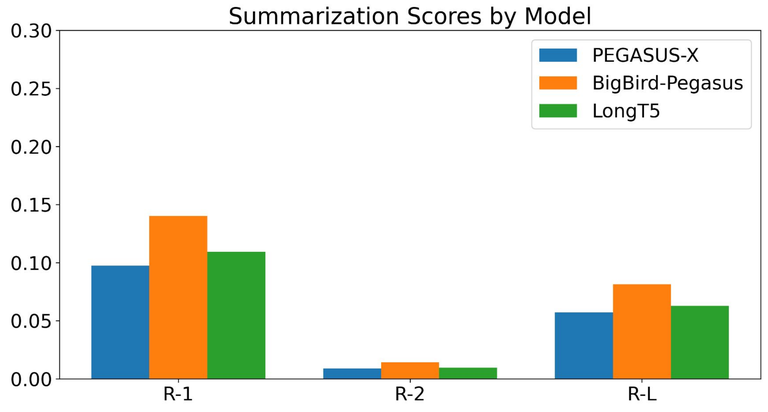

Figure 2. ROUGE scores for models compared on abstract summarization task

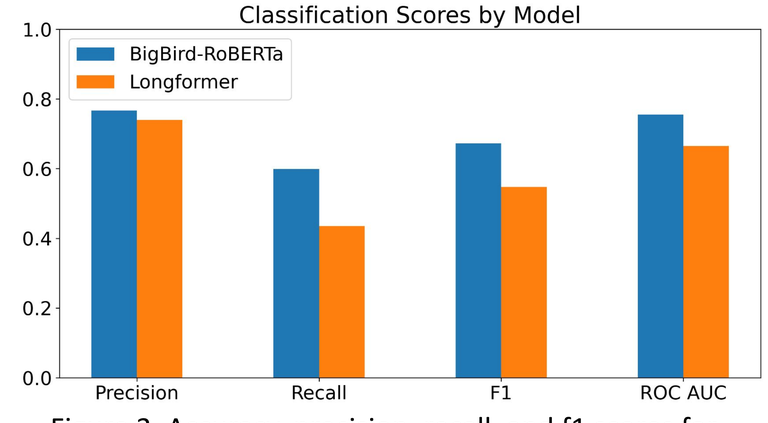

Figure 3. Accuracy, precision, recall, and f1 scores for classification approaches.

Discussion

Summarization:
- All pre-trained models showed very poor performance
- BigBird-Pegasus performed the best of the three models
- BigBird-Pegasus also had the best summaries upon human evaluation
- Future plans to fine-tune each model

Classification:
- BigBird-RoBERTa performed significantly better than Longformer by every metric
- Future plans to compare traditional machine learning approaches

Acknowledgements

Financial support was provided by the U.S. National Science Foundation through the Research Experiences for Undergraduates (REU) in AI/ML Deep Learning Accelerated Software-Hardware Co-Design under Dr. Hui Zhao and Dr. Mark V. Albert.