Leveraging Patient Portal Messages to Predict Emergency Department Visits

Jasmine Bilir, Tran Le, Julia Kadie

Department of Computer Science, Stanford University, School of Medicine

Problem

- **Stanford Medicine's goal:** reduce the number of preventable visits to the emergency department (ED)
- **Our research question:** Can using MyHealth patient portal message data improve prediction of 1-year risk of ED visit? (Binary prediction: ED Visit = 1)

Background

- Stanford Medicine currently uses a logistic regression model for ED prediction with F1 score of 0.396
- Stanford MyHealth is a patient portal allowing message exchange between patients and care teams
- Turchin et al. found that BioBERT and ClinicalBERT outperformed general BERT on medical concept recognition in outpatient provider notes (Alexander Turchin, 2023)
- Our cohort includes messages from adult patients seen by primary care between 2018/8/1 - 2020/8/1. A true label for ED visit means that a patient came to an ED in the next year.
- Our data composition includes:

Method	Train	Validation	Test
First-512 Bio_Clinical Bert Finetuning	Number of message rows: 2000 Number of patients: 2000 Prevalence rate: 21.55%	Number of message rows: 200 Number of patients: 28 Prevalence rate: 21.43%	Number of message rows: 200 Number of patients: 22 Prevalence rate: 18.18%
Chunked-512 Bio_Clinical Bert Finetuning	Number of message rows: 2000 Number of patients: 244 Prevalence rate: 19.67%	Number of message rows: 200 Number of patients: 28 Prevalence rate: 21.43%	Number of message rows: 200 Number of patients: 22 Prevalence rate: 18.18%
Experiments 1-4	NA	NA	Number of message rows: 2000 Number of patients: 187 Prevalence rate: 23.52%

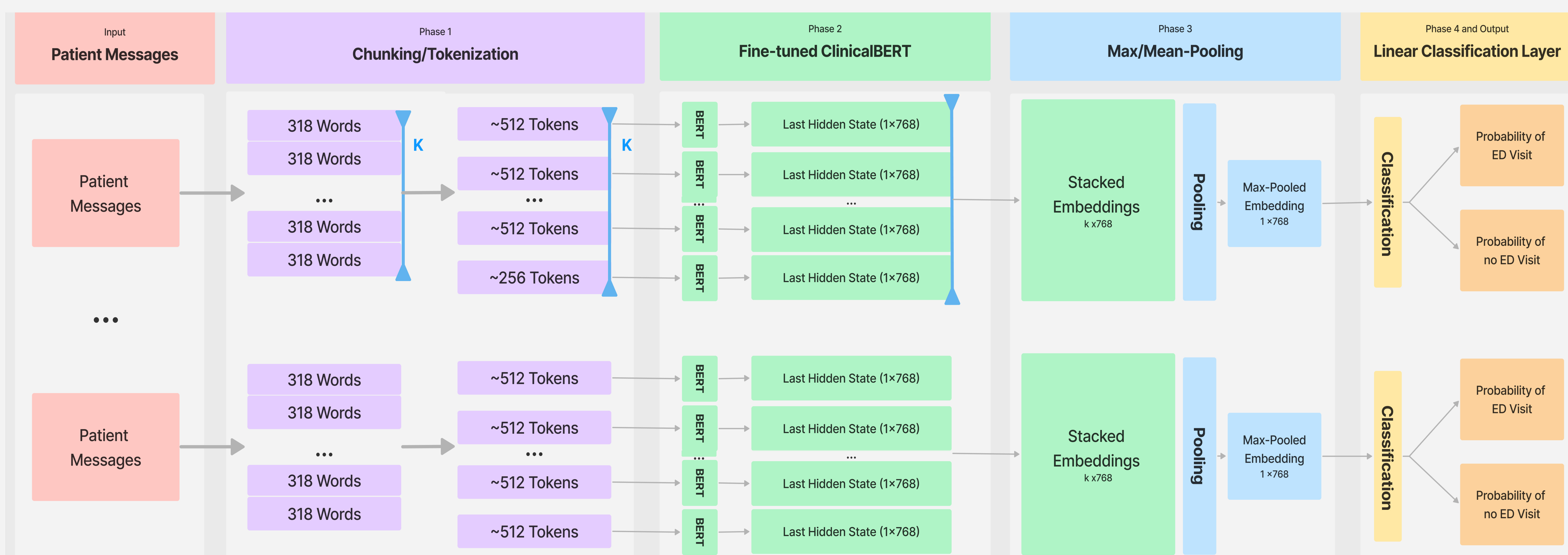
Methods

In approaching our problem, we trained models on the binary classification task of predicting ED visits in 3 major steps:

1. We fine-tuned the pretrained **Bio_Clinical BERT** model from **Huggingface** on our dataset. For training, We broke patient messages into chunks of size 512 and trained the model on 244 patient's messages, averaging 2740 words/patient.
2. To effectively use the content of an entire patient message history for our prediction task, we utilize the hierarchical approach of splitting each patient's message history into k chunks of size 318 words, approximately 512 WordPiece tokens. Patients had 9 chunks on average.
3. Next, we leverage our finetuned model and perform **4 post-processing experiments** to make this model adaptable to patient messages longer than 512

   These experiments included: **Max-Pooling**, **Mean-Pooling**, **Max-Voting,** and **Threshold of 1 Voting**

We present our pooling model below:



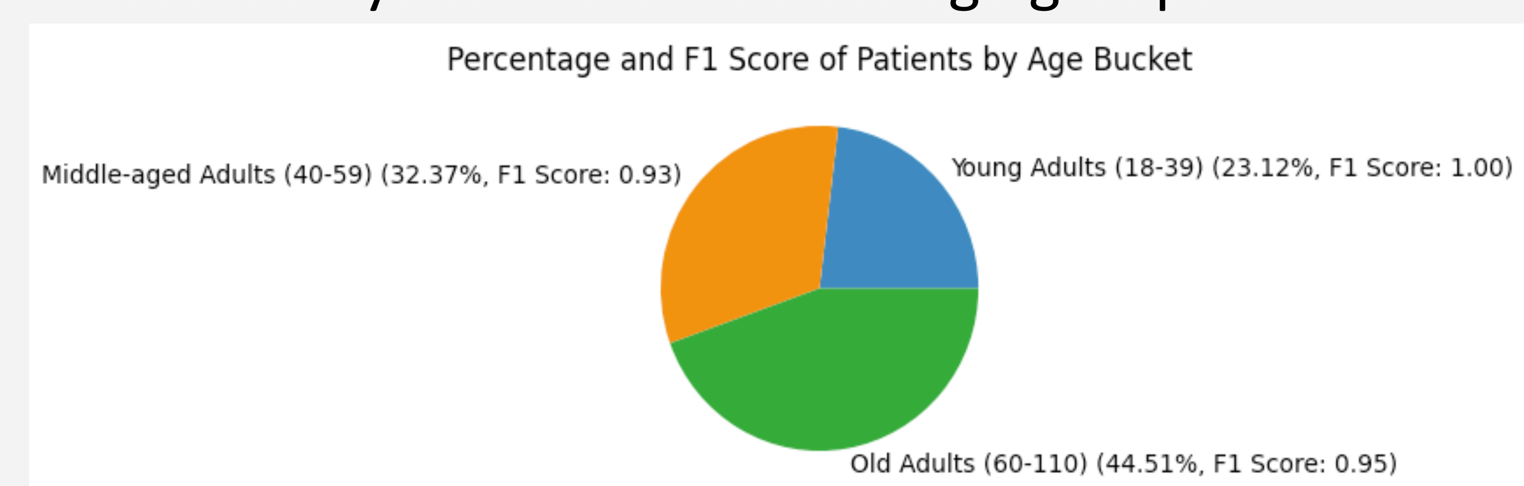
Experiments

- Our dataset contains three attributes: the patient id, patient messages text, and ED visit labels (T/F)
- We obtained the following results:

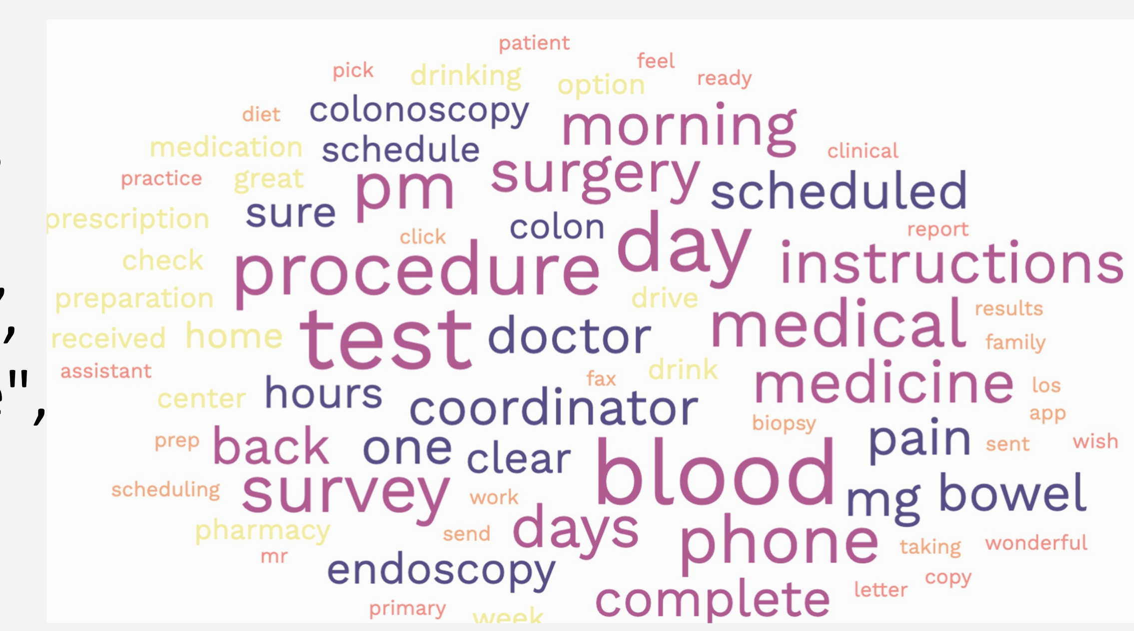
Model Configuration	Evaluation Metrics	Training Time
Epic Logistic Regression Model	AUROC: 0725 F1: 0.396	Unknown
First-512 Bio_Clinical BERT Baseline	AUROC: 0.563 F1: 0.213	6 min
Chunked-512 Bio_Clinical BERT + Max Pooling	AUROC: 0.990 F1: 0.941	1 hour 20 min
Chunked-512 Bio_Clinical BERT + Mean Pooling	AUROC: 0.993 **F1: 0.954**	1 hour 20 min
Chunked-512 Bio_Clinical BERT + Max Voting	AUROC: 0.997 F1: 0.938	1 hour 15 min
Chunked-512 Bio_Clinical BERT + Threshold Voting	AUROC: 0.985 F1: 0.733	1 hour 10 min

Analysis

- Mean pooling produces the best F1 score of 0.954
- Mean pooling reduces overfitting by averaging out noise in the input data and preserve the locality of the input data since we are chunking the message sequentially
- Both **pooling methods perform better than the voting methods**: voting simply considers whether there is a certain number of true predictions in the output, whereas pooling methods filter through all the inputs to make an informed prediction
- Baseline BERT method performs worst because only first 512 tokens of each message is considered, missing context
- F1 score stays consistent across age groups:



- **LDA analysis** on positive predictions show messages relating to "surgery", "blood", "procedure", "test", and "pain". See word cloud depiction:



Conclusion

- Our model **outperforms the current model** used by Stanford Healthcare for ED prediction
- Patient portal messages are useful in prediction of ED visits
- Next steps: Acquire more compute to run with larger datasets, combine our NLP work in larger model with patient attributes, conduct transparency analysis