Leveraging Patient Portal Messages to Predict Emergency Department Visits

Stanford CS224N {Custom} Project:
Jasmine Bilir, Julia Kadie, Tran Le
jbilir@stanford.edu, jkadie@stanford.edu, tranle@stanford.edu
Submitted using shared late-days: 6 from Tran Le, 3 from Jasmine Bilir / Julia Kadie

Abstract

Stanford Medicine's goal is to reduce the number of preventable visits to the emergency department (ED) due to many negative effects of ED congestion. This project's goal is to leverage robust NLP BERT models predict the probability that a patient will be admitted to the ED in the next year given their past medical messages to and from their care providers in the last two years. Our experiments yielded F1 scores of up to 0.954 and suggest that patient portal messages can play a vital role in predicting patient healthcare outcomes. With these novel approaches, we hope to contribute to the growing body of research using patient portal messages in NLP and show a definitive use case in ED visit prediction.

External collaborators: Dr. Julian Genkins, Stanford Medicine Physician and Clinical Informatics Fellow. Julian provided medical advice for the project. Hanna Kiani and Seth Cohen, Stanford Medicine MyHealth application team members who assisted with problem identification and data extraction.
Mentor: Abhinav Garg.
Sharing project: Not sharing with another class.

1 Introduction

1.1 Motivations

Stanford Medicine currently uses a logistic regression model V1 developed by Epic Systems Corporation that uses approximately 27 variables including medications, lab results, and more to predict patients who will be admitted to the hospital (Epic, 2016). The replicated Epic V1 model achieves an an F1 = 0.348 and AUROC = 0.710 on our experiment dataset. Therefore, the model is unable to accurately predict patients who will visit the ED and cannot easily scale to new factors that may be added to the patient's electronic health record (EHR) data. Epic's model shows the need for a more robust technical solution to decrease demands in the ED.

1.2 Our Contributions

At Stanford Medicine, the MyHealth application is used to communicate between the patient and care team. We hypothesize that leveraging patient portal messages between between patients and healthcare providers will produce more accurate predictions of patients who will be admitted to the ED. Specifically, we use a dataset from Professor Jonathan Chen's Health Rex Lab that includes Stanford Medicine patients, messages they send and receive, and whether or not they visit the ED. Our results show that finetuning a pretrained Bio_Clinical BERT model and mean pooling over the hidden states produce the best results with an F1 score of 0.954.

2 Related Work

Our literature review revealed the methods Stanford Medicine currently uses for the ED prediction task, pointed us to a proof of concept for using patient messages for ED visit prediction, identified the optimal pre-trained model to base our custom models on, and provided methods for dealing with clinical texts that are too long for the Bio_ClinicalBERT model.

The current model used by Stanford Medicine for the ED prediction task is a logistic regression model made by the private company Epic Systems. Its features are not NLP-related and do not include patient portal messages, but rather variables associated with patient health data such as "insurance status" and "presence/absence of diagnoses" (Epic (2016)).

The task of using patient portal messages for inference is being explored in different areas of medicine: Sulieman et al. found that patient portal messages can indicate risk of hospital readmission for patients with ischemic heart disease (et al (2020)). While this paper did not take a deep learning approach to identifying their findings, this work represents a proof of concept that patient portal messages may be used as indicators for a similar problem such as hospital readmission.

Previous work has shown promise in using large language models on narrative medical documents for inference tasks. Indeed, Turchin et al. found that BioBERT and ClinicalBERT outperformed general BERT on medical concept recognition in outpatient provider notes (Alexander Turchin (2023)). Although this dataset is not directly identical to ours, it is a proof of concept of how ClinicalBERT and BioBERT can be used on narrative-like clinical datasets. Thus, we used this paper to identify Bio_ClinicalBERT, a model combining the ideas of the 2 separate BERT models mentioned in this paper, as the foundation of our custom models (HuggingFace) .

Given the long nature of our message data, we also explored the limitations of using transformers on clinical documents. Gao et al. note that the long nature of clinical texts makes using transformers difficult given the limited input size of transformers ((et al., 2021)). With a pretrained BERT baseline model, they survey methods to resolve this size issue. First, they finetuned a pretrained BERT model for their classification task by only using the first 510 wordpiece tokens of an input. Then, they chunk clinical notes into k 510 wordpiece tokens, feed the chunks into the finetuned model, and and max pool over the k logits outputted by the finetuned BERT base model. Ultimately, this paper found that a non-transformer hierarchical self attention method was able to outperform the transformer methods. However, we hope to improve upon their findings by leveraging Bio_ClinicalBERT instead of BERT due to its domain knowledge in the bio-clinical sphere. While this paper implemented max pooling over logits, we will improve upon this by implementing max pooling over our last hidden states across our k chunks per patient to minimize information loss. Moreover, we experiment with other post processing such as mean pooling, max voting and threshold of one voting.

3 Approach

3.1 Tasks

Our task is to use all patient message data between 8/1/2018-8/1/2020 to make a prediction about whether or not (true or false) a patient will have an ED visit between 8/1/2020 to 8/1/2021. We then evaluate whether our approach was accurate using the labels in our dataset. A true label for ED visit means that a patient came to an ED and was (1) either discharged or (2) admitted to the hospital (normal floor or ICU).

3.2 Methods Overview

For our baselines, we used the existing Epic logistic regression model and also finetuned Bio_Clinical BERT on the first 512-tokens of each message in our dataset. For our advanced work, we finetuned Bio_Clinical BERT on all our messages by chunking our dataset into 512-token chunks. We then conducted four experiments where we added additional layers on top of our advanced fine-tuned Bio_Clinical BERT. Throughout all our experiments, we implemented document-chunking which was needed given the 512 max-token constraints of Bio_Clinical BERT.

3.3 Baselines

1. **Epic Logistic Regression:** For our first baseline, we reference the V1 Epic logistic regression model. The replicated Epic V1 model achieves an an F1 = 0.396 and AUROC = 0.725 on our experiment dataset (Epic, 2016).

2. **First-512 Bio_Clinical BERT:** Our second baseline utilizes the pre-trained Bio_Clinical BERT model which we finetune to our patient portal message dataset. We do this using only the first 318 words in the set of all patient messages for each patient. 318 words corresponds to 512 tokens, the max input size of the BERT model. Bio_ClinicalBERT was trained on a database containing electronic health records (EHR) data. The EHR data is a health medium nearly paralleling patient-clinician message data, but is optimized for shorter text sequences rather than the long set of messages we attempt to classify. For this reason, we finetune the model on the full message data by using chunking and pooling techniques to capture the full message text beyond the first 512 tokens.

Aside from using the pre-trained HuggingFace model and using Hugging Face API calls to finetune the model, all experiments were coded from scratch and are specific to this work.

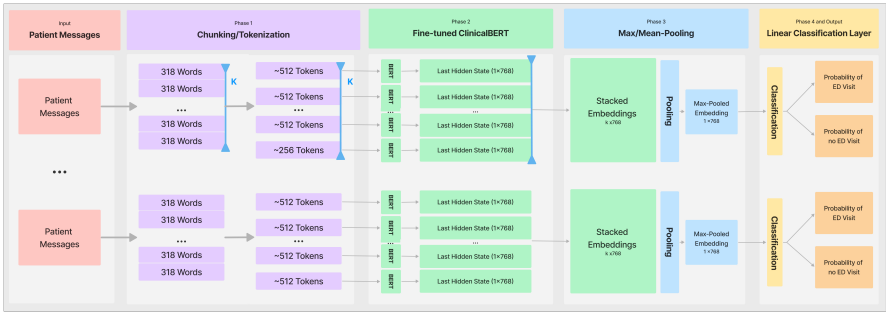
3.4 Chunked-512 Bio_Clinical BERT Finetune

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, which translates to approximately 512 wordpiece tokens. We then trained and saved the fine-tuned implementation of Bio_ClinicalBERT using all k message chunks (unlike our baseline which only used the first message chunk). In our experiments below, we then utilize various methods to extend this base architecture and produce one prediction for each patient.

3.4.1 Experiment 1: Chunked-512 Finetuned Bio_Clinical BERT + Max Pooling

Our finetuned Bio_ClinicalBERT for classification on each of the k segments generates k embedding vectors, one for each segment. This segment embedding represents the last hidden state of our pre-trained model. Before the classification step, we apply a max pool operation across all k embeddings to create one embedding vector representing the most relevant components of a patient's message history. This embedding vector is then passed through a linear classification layer to generate logits for the patient.

Figure 1: Experiments 1 and 2



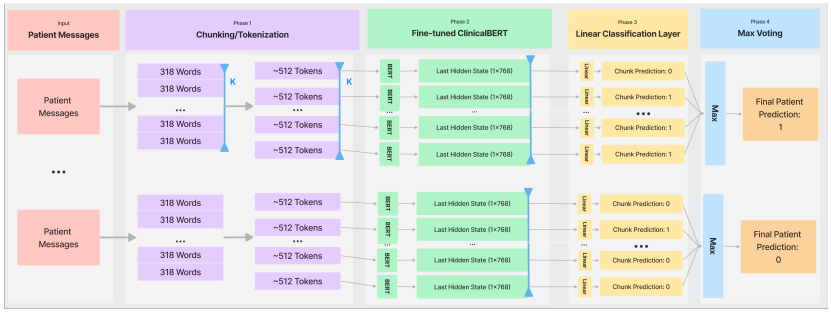
3.4.2 Experiment 2: Chunked-512 Finetuned Bio_Clinical BERT + Average Pooling

This experiment replicated the above max pooling experiment, but used the average values of the stacked embeddings for the classification task instead of the maximum values.

3.4.3 Experiment 3: Chunked-512 Finetuned Bio_Clinical BERT + Max Voting

This experiment utilized the fine-tuned Bio_Clinical BERT baseline to classify k individual message chunks independently. As in experiment 3, we used the fine-tuned Bio_Clinical BERT to produce k embedding vectors, one for each segment. Then, we pass these embeddings through the classification layer to produce k logit vectors for each of the k chunks, which we used to produce k predictions for each patient. Using these k predictions, we then found the prediction value that was the most common (either 1 or 0) across all k predictions, and assigned this value to be the final patient prediction.

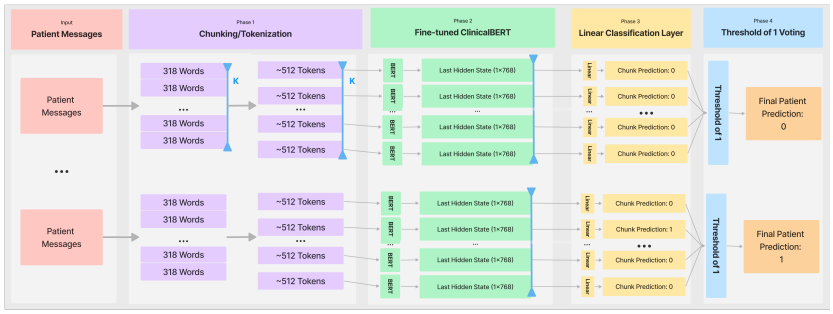
Figure 2: Experiment 3



3.4.4 Experiment 4: Chunked-512 Finetuned Bio_Clinical BERT + Threshold of 1 Voting

Because our task involves providing a medical prediction which could affect a patient and their well-being, we implemented an additional voting strategy that was more sensitive to any risk that might be apparent in a given message chunk. In this experiment we produced k predictions by following the sequence of steps in experiment 3. However, rather than take the max occurrence over predictions as in experiment 3, we used a threshold of 1. Put simply, if any of the k message chunks is labeled with prediction of 1, then we assign the overall patient prediction to be 1.

Figure 3: Experiment 4



4 Experiments

4.1 Data

Our dataset contains three attributes: the patient id, patient messages text, and ED visit labels (T/F) from patients using Stanford MyHealth. Our cohort includes adults seen by primary care between 2018/8/1 - 2020/8/1. We initially downsampled our base cohort to a 25k patient cohort with a 20% prevalence of ED visits. Then, this population was further downsampled in a non-random way to

patients who had at least one message sent or received in the 2 years prior to the prediction date so that we could train our model on texts. Overall, this population still looked similar based on high level demographic and disease characteristics.

In selecting our patient cohort, we only considered patients who were 18 and older (adults), alive at the prediction date, and had at least two encounters with Stanford primary care within the timeframe with meaningful visits with a provider, specifically: office visit or telemedicine visit. A true label for ED visit means that within 2020/8/1 - 2021/8/1, a patient came to an ED and was (1) either discharged or (2) admitted to the hospital (normal floor or ICU).

512 tokens is the maximum number of tokens that the BERT model will take. Thus, we pre-processed our data by splitting each patient's messages into chunks of 318 words. In our experiments, we found that 318 words approximately equates to 512 tokens. In our chunked dataset, each row represents 318 words from a patient message.

Due to GPU limitations, we selected a subsample of our messages as describe in the table. While the number of rows we used were consistent, the number of messages and prevalence varied.

Table 1: Data Composition of All Methods

Method	**Train**	**Validation**	**Test**
Baseline First-512 Bio_Clinical Bert Finetuning	Number of 512-token chunks: 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%

We run experiments 1-4 with a test set of 2000 rows because we wanted to verify our custom methods on a larger number of patients. Furthermore, the extensive computation power required to finetune the Bio_Clinical BERT model limited the number of test rows to 200 whereas the experiments did not have this limitation.

4.2 Evaluation method

The ultimate goal of predicting ED visits is for doctors to provide preventative care for at-risk patients. It is important that the model identifies patients at risk effectively, so that the medical system can provide them with necessary care to avoid the ED. On the other hand, due to limited resources in the medical system, it is important that the model does not over-risk patients, as that would divert important resources from those who need it most. As such, it is important to balance precision and recall. Thus, we evaluate our models using two main calculations. First, we look at the F1 score, which tells the model's balanced ability to both identify positive cases (recall) and be accurate with the cases it does capture (precision) where $F1 = \frac{2TP}{2TP+FP+FN}$. F1 score will be our main evaluation metric as the dataset is imbalanced.

Second, we calculate the AUROC score where a score of 1 means the classifier can perfectly distinguish between all the positive and the negative points. Let "true probability" represent the probability that our model predicts that a patient will be admitted to the ED (label of 1). For experiments 1 and 2, we obtain the logits from the output of the classifier. For experiments 3 and 4, since we did not perform pooling over the last hidden states, we select the logits for calculating the true probability. For experiment 3, we select the logits pair with the largest probability for predicting our max vote value for a given patient. For experiment 4, we select the logits pair with the largest probability for predicting 1 for a given patient. We then feed the logits in all four experiments into

the softmax function to calculate the true probability. The true probability is used to calculate the ROC AUC score.

4.3 Experimental details

Table 2: Bio_Clinical BERT Finetune Training Configurations

Method	**Model Configuration**	**Evaluation Metrics (Test Set)**	**Training Time**
Baseline First-512 Bio_Clinical Bert Finetuning	Epochs: 5 LR: 2e-5 Metric for best model: F1 Train set: 2000 rows	Loss: 1.15 AUROC: 0.678 F1: 0.316	6m
Chunked-512 Bio_Clinical Bert Finetuning	Epochs: 6 LR: 2e-5 Metric for best model: F1 Train set: 5000 rows Gradient accumulation steps=8 fp16=True	Out of GPU memory	NA
Chunked-512 Bio_Clinical Bert Finetuning	Epochs: 5 LR: 2e-5 Metric for best model: F1 Train set: 2000 rows	Loss: 0.201 AUROC: 0.987 F1: 0.897	1 hour

4.4 Results

Table 3: Quantitative Results on Experiments

Model Configuration	**Evaluation Metrics (Test Set)**	**Experiment Time**
Epic Logistic Regression Model	AUROC: 0.725 F1: 0.396	Unknown
Baseline 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	20 min
Chunked-512 Bio_Clinical BERT + Mean Pooling	AUROC: 0.993 **F1: 0.954**	20 min
Chunked-512 Bio_Clinical BERT + Max Voting	AUROC: 0.997 F1: 0.938	15 min
Chunked-512 Bio_Clinical BERT + Threshold Voting	AUROC: 0.985 F1: 0.733	10 min

From the evaluation metrics, we see that the mean pooling experiment produces the best F1 score of 0.954 while the BERT baseline produces the worst F1 score of 0.213. This is what we expected because the BERT baseline only takes in the first 512 tokens of each messages, which means that many words from the messages will not be used for the final prediction. We also see that both pooling methods do better than the voting methods. This makes sense since max and threshold voting simply consider whether there is a certain number of true predictions in the output, whereas the pooling methods filter through all the inputs to make an informed prediction. Lastly, we see that mean pooling performs better than the max pooling method. We attribute this to mean pooling's ability to reduce overfitting by averaging out noise in the input data and preserve the locality of the input data since we are chunking the message sequentially. Overall, mean pooling's ability to capture important features distributed across different chunks of the messages results in the most accurate predictions.

5 Analysis

5.1 Manual Qualitative Analysis of False Negatives and True Positives

Due to the narrative nature of our patient message data, we manually read our true positive examples to identify trends. We noticed a few key trends:
(1) Messages indicated that patients had upcoming procedures and included lots of instruction text about those procedures. However, we cross-checked with our false negative examples and also saw false negatives referring to procedures, so we conclude that those features are not defining to our model for the prediction task. Additionally, the two procedures mentioned in the false positive cases are also included in the true positive cases, so the *type* of procedure learned by the model does not contribute to a difference in predictions.
(2) Patients in the set of true positives often describe in detail the pain that they are going through to their providers. From a qualitative perspective, it is easy to see that patients have deep concerns for their health. As an anonymized example: *"my hand was going numb and I had shooting nerve pain in my fingers...I'm concerned about nerve damage"*. Additionally, many of the patients are in conversation with their care providers about pain management of their conditions, signifying that they have conditions or have had recent procedures. These common patterns suggest that the model identifies the sentiment of pain levels and factors this into a positive prediction.

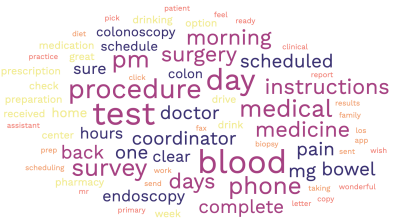
Additionally, we manually read all of our 3 false negative examples in the test set. In each false negative example, the entire patient message history only included messages sent to the patient - each false negative patient never responded to a message. This could signify that our model requires data from patients to accurately infer ED predictions and that patient-provider message exchanges are the key piece that our model learns. This hypothesis is supported by analysis of true positives. Indeed, after examining our 31 true positive patients, only $1/31$ had a set of messages in which they never responded to their provider. For example, an anonymized and shortened patient message history of a false negative example is shown to the right where the patient never responded.

Dear Mr. [Name],
Message
Sincerely,
Stanford Primary Care Physician
...
Dear Mr. [Name],
Message
Sincerely,
Dr. [Name2]
...
Dear Mr. [Name],
Message
Sincerely,
Dr. [Name2]

5.2 Latent Dirichlet Allocation Topic Modeling

Additionally, we implemented and ran Latent Dirichlet Allocation Topic Modeling (LDA) to better interpret the model's classifications. To do this, we ran LDA separately on positive and negative predictions, and generated the top 10 topics for each. These were not necessarily correct predictions, but we chose to use all predictions in order to gather information on the patterns the model used to make distinctions. The two sets of 10 topics were then used to generate the subsequent word clouds where the word size is represented by their occurrence level.

Figure 4: LDA WordCloud for ED Prediction



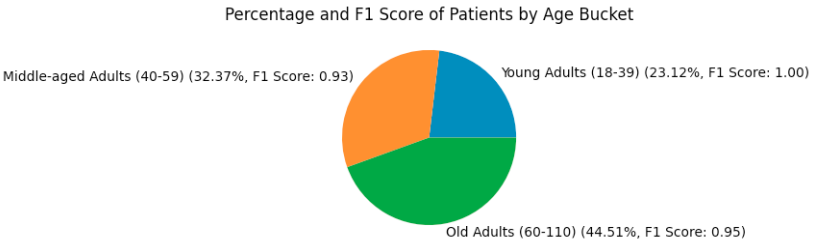
In 4, the figures found for positive attributions showed repeated instances of "procedure", "test", "instructions", "blood", "surgery", and "pain". Depending on when and how these terms appeared in the patient's message history, it is possible that the model picked up on worsening health conditions or medical complications of the patient. Other words that appeared but were not as common in the topic analysis were "morning" and "prescriptions". Possible hypotheses are that patients might reach out to providers seeking help in the morning if they are dealing with pain and considering visiting an

ED. Likewise, patients seeking immediate help getting a prescription might later go to an ED if they are unable to get support from their primary doctor.

Figure 5: LDA WordCloud for No ED Prediction

As for negative predictions in 5, we saw common words appear in topic analysis such as "day", "need", "call", "health", "instruction", "questions", "procedure", and "hours". Our takeaway from this information is that the model may have used messages that were more logistical (involving scheduling, pre-care/post-care instructions, etc) to indicate decreased risk of having an ED visit. There are words that appeared in this set of topic as in the first such as "blood", "surgery", and "procedure". We are unsure what decisions the model made to differentiate these cases; however, refering back to our manual analysis it is possible the model did not view these words as distinguishing factors. It may also be that Bio_ClinicalBERT used message context to understand distinctions between these types of messages. Another hypothesis is that these messages may have been resolved and well managed at some point during the patient's message history which put the patient at decreased risk. Other words that appeared in the topic modeling, although less frequently, were "good", "healthy", "young", "muscles", and "complete". We interpret these as indications of good health and show that the model may have been correlating such terms with decreased ED visit risk.

5.3 Evaluation Metrics Across Demographics



Lastly, our data analysis shows that our model is producing accurate predictions across different types of patients. The figure above represents the composition of the test set predictions for mean pooling and the respected F1 score for each group. We see that mean pooling produces high F1 scores of 1.0, 0.93, and 0.95 for young, middle-aged, and old adults respectively.

6 Conclusion

In this research project, we used patient portal messages to predict future patient ED visits. We learned that applying mean pooling on a finetuned pretrained Bio_Clincal BERT model achieved high F1 and AUROC scores for patients across different age demographics. By analyzing our model, we learned that certain words when contextualized in patient messages like "surgery", "pain", and "test" might indicate increased risk for an ED visit. Likewise, there might be cases where words common to logistical questions and consistent care might indicate lower ED risk. Since our model performs better than the current Epic V1 logistic regression model used by Stanford Medicine, we hope to work with doctors at Stanford Medicine to further develop and implement our work. Specifically, we would like to increase our GPU resources to run on more data, sample different cohorts, and test our methods to ensure it is not biased against patients from different demographics.

References

Lina Sulieman et al. 2020. Why patient portal messages indicate risk of readmission for patients with ischemic heart disease. In *AMIA Annual Symposium Proceedings Archive*, online.

Shang Gao et al. 2021. Limitations of transformers on clinical text classification. In *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS*, online.

Stanislav Masharsky et al. Alexander Turchin. 2023. Comparison of bert implementations for natural language processing of narrative medical documents. In *Informatics in Medicine Unlocked*, online.

Epic. 2016. Cognitive computing model brief: Hospital admissions and ed visits version 1. 1979 Milky Way. Epic Systems Corporation.

HuggingFace. Bioclinicalbert.