Cover Page

Project Name: Binary Classification of Sepsis Development Using LLM Technology

Prepared By: Sampath Martha, Nick Smith, Joseph Walker

Data: April 29th , 2024

Presentation Link:
https://docs.google.com/presentation/d/1OEeqc_chuDNamWv248IJ4bwzflf1CHNzDs5w_dDnlaw/edit?usp=drive_link

Code Path:
https://drive.google.com/drive/folders/1k4PQRN3nuofTk7_xAKYwgaS4akea0EQ1?usp=drive_link

Presentation Video:
https://utexas.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=dac553a3-c990-4eb4-acbf-b162004fd851

Or

https://drive.google.com/file/d/1fdcu-KlqNL5d8R5TqOZHfnUce7Kx921u/view?usp=drive_link

Binary Classification of Sepsis Development Using LLM Technology

Sampath Martha
Computer & Data Science
Online
University of Texas at Austin
Austin Texas United States
itsampathm@gmail.com

Nicholas Smith
Computer & Data Science
Online
University of Texas at Austin
Austin Texas United States
nicholassmith@utexas.edu

Joseph Walker
Computer & Data Science
Online
University of Texas at Austin
Austin Texas United States
josephwalker.utexas@gmaill.com

1 Introduction

According to the Centers for Disease Control and Prevention (CDC), Sepsis is “the body’s extreme response to an infection. It is a life threatening medical emergency [1].” Every year at least 1.7 million adults will develop sepsis, and at least 350,000 of adults who develop sepsis die from it. Sepsis itself is not an infection but is the body’s response from infection, therefore sepsis can develop from various infections, often caused by bacterial, viral, or fungal infections. This also means that sepsis itself is not contagious.

Sepsis is considered to be a life-threatening medical emergency and patients must receive immediate care. Early detection of sepsis can be critical for taking preventative or intervention measures to save a patient's life [4]. In this study, we will focus on the use of MIMIC-III electronic health records (EHRs) to fine tune the LLM. The goal of this study is to use a fine-tuned Large Language Model (LLM) for detection of sepsis in patients and compare its performance to Machine Learning (ML) models trained on raw last-layer embeddings of another LLM. Comparison of results will also be provided against a real-world screening tool used in clinical practice as a performance benchmark.

2 Related Work

Houston Methodist Hospital put into place a sepsis screening tool conducted by nurses on staff to detect whether a patient has sepsis. They found employing their tool led to earlier detection of sepsis, and thus a reduction of sepsis mortality. Their tool was evaluated to have “a probability that sepsis was present in a patient who screened positive (positive predictive value) was 80.2%, while the probability that sepsis was absent when a screen was negative (negative predictive value) was 99.5% [4].”

The National Library of Medicine performed a literature review of ML models that attempted to detect sepsis [3]. identified when performing a study on sepsis, there are several important aspects to keep in mind. Firstly, the sample size should be large enough to ensure that the results are statistically significant. Secondly, the data should be accessible and available to other researchers in order to ensure transparency and replicability. Thirdly, it is important to consider how the model will be used in the real world. If the goal is to develop a sepsis screening tool for use by healthcare professionals, then the model should be designed in a way that is easy to use and interpret. It is also important to consider the potential for bias in the data. Missing data can be a significant problem, and it is important to have a strategy in place for dealing with missing data in order to maintain data integrity. It is important to use a suitable machine learning model and to carefully tune the hyperparameters in order to prevent overfitting. The results of the study should be clearly reported, and it is important to discuss both the clinical implications of the findings and the limitations of the study.

A fine-tuned LLM, using tabular data, was demonstrated to outperform state-of-the-art ML models when smaller datasets are available and were comparable when the datasets were larger [2].

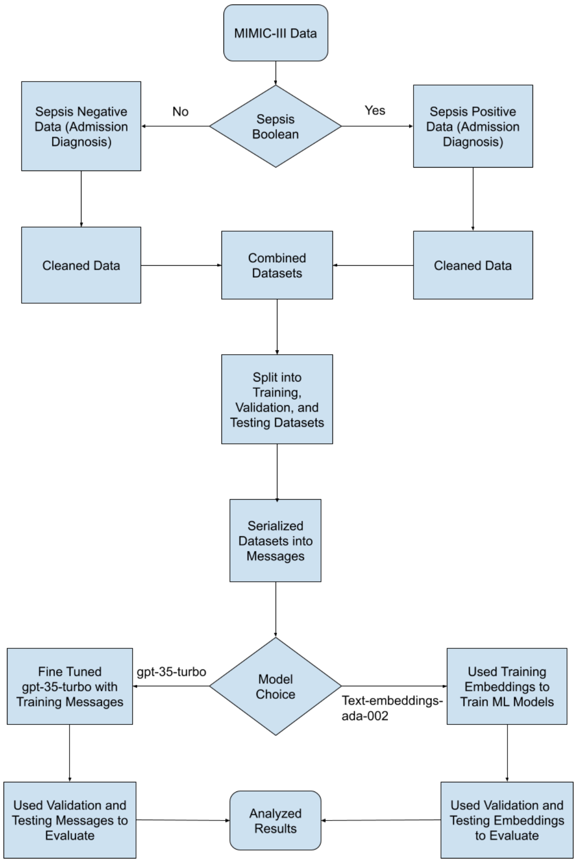

MIMIC-III Data
Sepsis Boolean
No
Yes
Sepsis Negative Data (Admission Diagnosis)
Sepsis Positive Data (Admission Diagnosis)
Cleaned Data
Combined Datasets
Cleaned Data
Split into Training, Validation, and Testing Datasets
Serialized Datasets into Messages
Model Choice
gpt-35-turbo
Text-embeddings-ada-002
Fine Tuned gpt-35-turbo with Training Messages
Used Training Embeddings to Train ML Models
Used Validation and Testing Messages to Evaluate
Analyzed Results
Used Validation and Testing Embeddings to Evaluate

Figure 1: **Process map of workflow from raw data to results.**

3 Methodology

Please see Figure 1.

3.1 Significant Features

The top-significant features of patients diagnosed with sepsis were first researched. There is an at risk set of people such as adults age 65 and older or those with chronic medical conditions or compromised immune systems. For use of MIMIC-III we had to research relevant metadata and down-select to the top-significant features. The definition for sepsis has developed over time and the two commonly accepted definitions are Sepsis-2 and Sepsis-3. Sepsis-3 focuses on potentially fatal organ malfunction or failure. Sepsis-2 is focused on systemic inflammatory response syndrome (SIRS). The Sepsis-2 definition provided the target features for this project. This target feature set became age, white blood cell count (WBC), body temperature, heart rate, respiratory rate [3]. It was deemed appropriate to use Sepsis-2 definition because this was the standard definition from 2001 to 2016 and the MIMIC-III data was populated within that time frame.

3.2 Cleaning the Data

With the target features in mind, the next step was to use SQL to query multiple MIMIC-III files to retrieve the required data. Our initial query discovered 1096 patients that were diagnosed with sepsis at admission. After making multiple considerations on the data we retrieved, which is discussed in section 3.5, we ended up refining the data. Cleaning data criteria included: making sure all ages were between 65 and 120; all positive patients later received sepsis-positive icd9-codes and all negative patients did not receive sepsis-positive icd9-codes; all feature measurements were within 24 hours of admission and not duplicate; and that there were no missing feature measurements. After cleaning, a positive sepsis dataset of 258 patients remained. For the negative sepsis dataset, the dataset was randomly sampled to a balanced 258 patients. Both datasets were combined for a total of 516 patients. Finally, all patients were shuffled and split 70/15/15 for training, validation, and testing datasets.

3.3 Serializing the Datasets

Once the DataFrames were set the next step is to serialize the data into a format that can be read by the LLM. In this case we used gpt-35-turbo_0613. Before combining the positive and negative datasets into a single dataframe, an extra column was added to each as a boolean for if the patient is part of the positive or negative dataset. This would be used to tell the model if the patient should be positive or negative during training. This would also allow us to double check during the validation process. For fine tuning, the process required the data be serialized into a jsonl format, so we created an additional script to deserialize the csv, convert all of the patient data into column prompts, and then serialize into the jJSONL format.

3.4 Model Choice

Now that the data was converted into a GPT legible format, the next step is to train the model. We tested on multiple models so it was important to pass the same prompts to each model. We followed the format created in the TabLLM Few-shot report and after doing some prompt engineering our prompt became “Does this patient most likely have sepsis? Reply only with a Yes or a No as your answer.” This prompt was used for all models to make sure we have a direct comparison.

Initially, we fine tuned the gpt-35-turbo_0613 model accessed through Azure OpenAI to create the SepsisGPT model deployment. The hyper parameters used for fine-tuning were n_epochs = 7, batch_size = 1, learning_rate_multiplier = 0.1. We then generated last-layer text embeddings from the text-embeddings-ada-002 model (also accessed through Azure OpenAI) for all datasets: training, validation, and testing. These embeddings were applied to ML algorithms, including LogisticRegression, DecisionTreeClassifier, RandomForestClassifier, SVC, and XGBClassifier.

3.5 Considerations

Multiple changes were made over the course of the project. One consistent update made was modification of the DataFrames and target features. Initially, we took the approach of using more data than just the target features with the thought that the LLM could potentially find a pattern with more information. The initial DataFrame included features such as diastolic and systolic arterial blood pressure and the invasive blood pressure mean, among other features. We found that these extra columns became unnecessary so they were trimmed from the original dataset. From running tests on the LLM it was determined that the target features actually required were Heart Rate, Respiratory Rate, Body Temperature, and White Blood Cell Count.

We further found that there were a ton of patients with NaN values in the dataset which left concerns for training an LLM. Because we're using an external LLM which essentially operates like a black box, we weren't sure how the LLM would handle NaN values. When running initial tests with NaN data the LLM would either operate normally by making assumptions, return an error on the NaN data, or operate by using other data instead. Because of this unpredictable behavior we elected to remove patients with NaN values in the target features.

Another factor was the timeframe of the data. The data was filtered to be within a 24-hour time window to ensure that the feature measurements correlated to the state the patient was in [4]. For the negative patients the data was filtered out for sepsis to ensure that patients were targeted that never developed sepsis later in time. The team took multiple passes of cleaning the dataset to remove data not conforming to criteria listed in 3.2. This further reduced the dataset and we were left with a clean dataset of 258 patients positive for Sepsis and 258 patients negative for sepsis.

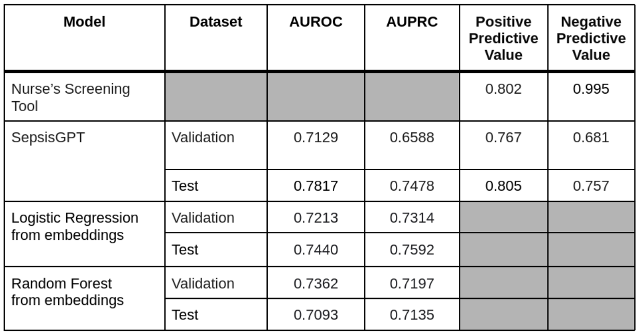
Another major consideration was the sample sizes used for training the model. The sample size factor affects the runtime of training the data, leading to hours of training on a single dataset. Additionally the sample size can affect the effectiveness of a trained model.

4 Results

We were able to successfully create a fine-tuned model, which we called SepsisGPT, by using the fine-tuning process. While we hoped for a higher performance score, we found it was comparable and even performing better than other algorithms such as Logistic Regression and Random Forest.

4.1 AUROC / AUPRC Scores

In Figure 2, you can see the performance of the trained models, and we use AUROC and AUPRC ratings to evaluate the performance of binary classification models. The nurse's screening tool is also provided as a benchmark value that was achieved in a clinical setting.



Model	Dataset	AUROC	AUPRC	Positive Predictive Value	Negative Predictive Value
Nurse's Screening Tool				0.802	0.995
SepsisGPT	Validation	0.7129	0.6588	0.767	0.681
	Test	0.7817	0.7478	0.805	0.757
Logistic Regression from embeddings	Validation	0.7213	0.7314		
	Test	0.7440	0.7592		
Random Forest from embeddings	Validation	0.7362	0.7197		
	Test	0.7093	0.7135		

Figure 2: **Model comparison and benchmark values for the nurse's screening tool.**

5 Conclusion

5.1 Summary

LLMs can be used to enhance the detection of sepsis in patients, which is critical for prevention or intervention of this life-threatening medical emergency. In this study we trained algorithms to detect sepsis in patients. In this study we were able to compare the performance of multiple Machine Learning algorithms in the detection of sepsis as discussed in section 4.

By taking MIMIC-III data and applying filters, we're able to select a group of patients based on a target feature set. After taking the selection of patients, we're able to train the algorithms on two DataFrames, the first containing patients positive with sepsis, and the second containing patients that never develop sepsis. After training the algorithms we were able to compare the performance based on the AUROC and AUPRC ratings. From this we are able to determine that the SepsisGPT model trained with fine-tuning performed the best.

5.2 Future Directions

There are a few improvements we focused on discussing. The first is the featureset. We only have 4 target features. However, from reading reports it seems that the focus is on abnormal values. To show abnormal values per patient it would be helpful to use a patient's historical data (baseline values) so then we can highlight abnormalities based on the patient instead of abnormalities on the average set of patients.

Additionally we were using the Sepsis-2 definition instead of the newer Sepsis-3 definition. The Sepsis-3 definition was created to “increase the precision and clinical utility of sepsis detection [3].” This definition simplifies by focusing on organ dysfunction rather than the target features specified in the Sepsis-2 definition, leading to less false-positive classifications that can be found using Sepsis-2. Lastly, the Sepsis-3 definition has set the standard for sepsis diagnosis and research which would allow this study to be more comparable to other studies.

Further improvements could also include evaluating the model performance using data from patients of all ages, experimenting SepsisGPT with various hyperparameter values (lower learning rate, more epochs, etc.), and designing the LLM model to predict the sepsis disease progression based on the temporal data.

REFERENCES

[1] *What is sepsis?* (2023) *Centers for Disease Control and Prevention*. Available at: https://www.cdc.gov/sepsis/what-is-sepsis.html (Accessed: 30 April 2024).

[2] Hegselmann, S. *et al.* (2023) *Tabllm: Few-shot classification of tabular data with large language models*, *arXiv.org*. Available at: https://arxiv.org/abs/2210.10723 (Accessed: 30 April 2024).

[3] Islam, K.R. *et al.* (2023) *Machine learning-based early prediction of sepsis using electronic health records: A systematic review*, *Journal of clinical medicine*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10488449/ (Accessed: 30 April 2024).

[4] Jones SL; Ashton CM; Kiehne L; Gigliotti E; Bell-Gordon C; Disbot M;Masud F; Shirkey BA; Wray NP; (2015) *Reductions in sepsis mortality and costs after design and implementation of a nurse-based early recognition and response program*, *Joint Commission journal on quality and patient safety*. Available at: https://pubmed.ncbi.nlm.nih.gov/26484679/ (Accessed: 30 April 2024).