CS598 DL4H in Spring 2022

Joseph Allen and Sonal Sharma
jaallen2@illinois.edu sonals3@illinois.edu

Group ID: 121, **Paper ID:** 215
Presentation link: TODO
Code link: https://github.com/josepha1/DLH_Final_Project

Introduction

Our team is trying to replicate **Contextual Embeddings from Clinical Notes Improves Prediction of Sepsis** by Fatemeh Amrollahi, Supreeth P. Shashikumar, Fereshteh Razmi and Shamim Nemati for our Final Project.

Sepsis, a life-threatening organ dysfunction, is a clinical syndrome triggered by acute infection and affects over 1 million Americans every year. Early detection of sepsis and timely antibiotics administration is known to save lives. In this work the authors have designed a sepsis prediction algorithm based on data from electronic health records (EHR) using a deep learning approach. The findings indicate that incorporation of clinical text features via a pre-trained language representation model (ClinicalBert in this case) can improve early prediction of sepsis and reduce false alarms.

The proposed method replaces the traditional tf-idf features with contextual embedded representations learned using Clinical-BERT, a state-of-the-art model for word and document embedding and specifically trained on a corpus of Biomedical and clinical texts. The resulting document-level representations were then concatenated with physiological, laboratory, and demographic information to make predictions of sepsis using a long-short term memory (LSTM) recurrent neural network.

Scope of reproducibility

Contextual word embedding models such as ELMo and BERT have dramatically improved performance for many natural language processing (NLP) tasks. However, these models have been minimally explored in specialty corpora, such as clinical text; moreover, in the clinical domain, no publicly available pre-trained BERT models yet exist. ClinicalBert addresses this need by exploring and releasing BERT models for clinical text.
The paper claims that using a domain-specific model with LSTM for prediction yields performance improvements for sepsis prediction.

Addressed claims from the original paper

- Combining both structural clinical data with ClinicalBERT embeddings (Model IV) achieved the best AUC performance.
- The ClinicalBERT approach outperformed the TF-IDF model for extraction of features from clinical texts.
- ClinicalBERT generated more meaningful representations of clinical notes and enabled the model to predict onset of sepsis more accurately.

Methodology

We tried to re-implement the approach from the paper description as the author's code wasn't accessible to us. Publicly available MIMIC-III dataset was used to reproduce results similar to the paper.

3.1 Model descriptions

The paper uses ClinicalBERT and LSTM models for sepsis prediction and performance improvements.

ClinicalBert learns deep representations of clinical text using two unsupervised language modeling tasks: masked language modeling and next sentence prediction. A clinical note input to ClinicalBert is represented as a collection of tokens. In ClinicalBert, a token in a clinical note is computed as the sum of the token embedding, a learned segment embedding, and a position embedding. ClinicalBERT improves over BERT on the MIMIC-III corpus of clinical notes.

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single data points, but also entire sequences of data.

3.2 Data descriptions

Publicly available MIMIC-III (Medical information for intensive care) dataset of critically ill patients, which includes anonymized physiological and clinical data, as well as clinical notes from over 50,000 intensive care unit (ICU) admissions collected between 2001 and 2012 was used as part of the study. Data includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more.

MIMIC-III is a relational database consisting of 26 tables. Five tables are used to define and track patient stays: ADMISSIONS; PATIENTS; ICU STAYS; SERVICES; and TRANSFERS. Another five tables are dictionaries for cross-referencing codes against their respective definitions: D_CPT; D_ICD_DIAGNOSES; D_ICD_PROCEDURES; D_ITEMS; and D_LABITEMS.

Access to the PhysioNet database was requested and obtained following the steps mentioned in piazza.

3.3 Hyperparameters

TODO

3.4 Implementation

Our team decided against using existing code and attempted to replicate the results of the authors by implementing the code on our own. In order to properly analyze the clinical notes contained in the dataset, preprocessing helper functions needed to be implemented. These helper functions were responsible for removing any unnecessary punctuation or characters that would interfere with the prediction process. An additional purpose of these functions was to decompose the clinical notes into sentences and limit the size of the resulting sentences so that they could be fed to the ClinicalBERT model for processing. Our team then loaded the data tables that contained the clinical notes and the ICD9 diagnosis codes. These tables were

combined so that the true labels for the diagnoses could be identified and later compared to the predicted results. The columns that were not necessary for our analysis were then dropped from the combined table and a final table was created for use with the ClinicalBERT model. To conduct our initial analysis and the efficacy of our code implementation, a sample of 89 data points were taken from the final table. Our team felt that this was a sufficient sample size to obtain meaningful results. The preprocessed sentences were then tokenized and input into the ClinicalBERT model. Each processed sentence was then concatenated with the other processed sentences from each clinical note to create a feature vector. The mean of each of these feature vectors were then input into the Long-Short-Term-Memory (LSTM) neural network. The output of the neural network was finally input into a SoftMax activation layer to get the final probabilities. Our team decided to use a probability threshold of 0.5 (50%) for our sepsis predictions. Our team decided to use this threshold in order to be consistent with the implantation of the original paper.

3.5 Computational requirements

For the purposes of our initial analysis, the computational requirements were quite minimal. The code was implemented on a Dell XPS15 with 32 GB of RAM and an 11th Gen Intel® Core™ i7-11800H Processor operating at 2.30 GHz. The average time it took to load all the data was approximately 1 minute and 23 seconds and the average runtime for analyzing the subset of 100 data points was approximate 1 minute and 11 seconds. The full dataset is approximately 4.2 GB, which is a moderately large dataset, but also significantly larger than the subset of data used for the initial analysis. Preprocessing all of the data, inputting it into the ClinicalBERT model, as well as training the LSTM, will take significantly longer when the complete dataset is used, however, our team feels as though the computational requirements for this task will easily be met.

4 Results

After inputting the sampled data into the ClinicalBERT model and LSTM network, our team obtained an **accuracy of 98.9% and an AUC-ROC score of 0.5.** While the accuracy of the model may indicate that the classifier is by and large predicting the correct sepsis labels, the AUC-ROC score indicates that the model is not distinguishing between classes very well. A possible explanation for the deviation of results from the original paper is that our team did not concatenate the results from the ClinicalBERT model is the physiological, laboratory, and demographics information utilized by the authors. One of the original claims in the author's work was that "Combining both structural clinical data with ClinicalBERT embedding (Model IV) achieved the best AUC performance", which seems to validate why there is a discrepancy between our accuracy results and AUC-ROC score results. Our team intends to further explore this claim by implementing the process followed by the authors in more detail. We believe that this will allows us to achieve results similar to the original paper, and therefore validating their results.

4.1 Result 1
4.2 Result 2
4.3 Additional results not present in the original paper

5 Discussion

TODO

References

TODO