CS230

Predicting Severe Sepsis from Electronic Health Records using Multi-Modal Deep Learning

Gowri Nayar
Massimo Giordano
Department of Computer Science
Stanford University

Abstract

Patients that will develop sepsis often present within the Emergency Department with a quick progression of the infection. In order to identify patients at risk for sepsis, the common methodology relies on the calculation of a score on the vitals, that depends on vitals measurements that are taken as the patient remains in the hospital. However, this score is not always accurate and are not representative of the full medical knowledge that is contained within the clinical notes. Therefore, we investigate a method based on the classification of triage notes, that can identify the risk for developing sepsis, in addition to the vital measurements. From this analysis, we can see the vitals measurements do outperform the clinical notes, and even in combined use, the clinical notes adds significant noise to the model.

1 Introduction

Overcrowding of emergency departments causes a decrease to patient care and an increase in mortality and hospital expense.[1] Data that is available when a patient presents to the Emergency Room, usually through the triage process, can be used to make predictions about the patients' needs and eventual hospital admission.[2][3] This study aims to use natural language processing on nurses' triage text documentation, which are crucial in understanding the patients' status, in combination with classification methods on the vitals measurements of the patients. By combining these two different modes of learning, we aim to create a stronger predictor in determining the outcome of the patient (sepsis, septic shock or discharge) and can allow for earlier clinical intervention to prevent adverse outcomes. There have been previous studies in implementing learning methods on the clinical notes or the vital measurements that are included in Electronic Health Records (EHR) data, but minimal work in combining such methods. ClinicalBERT[4][5] is the state of the art tool published that relies on an embedding on of words used in clinical settings. This model has been published and many Natural Language Processing (NLP) tasks that are implemented in clinical settings are trained using this tool. From the Targetted Real-time Early Warning System [6] implemented at Johns Hopkins ICU, we can see the impact of using vitals measurements in EHR data for the classification task. We aim to draw inspiration from these two methods to create a multi-modal learning model that combines the predictive power of the text and numerical data. This work focuses on comparing the performance of two models trained on the text and numerical data, and a third model that trains the two together.

2 Related work

This work draw inspiration from a recent publication by Sterling et al. who describe a methodology to predict emergency department outcomes using NLP methods on triage notes. [7] This paper reports

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

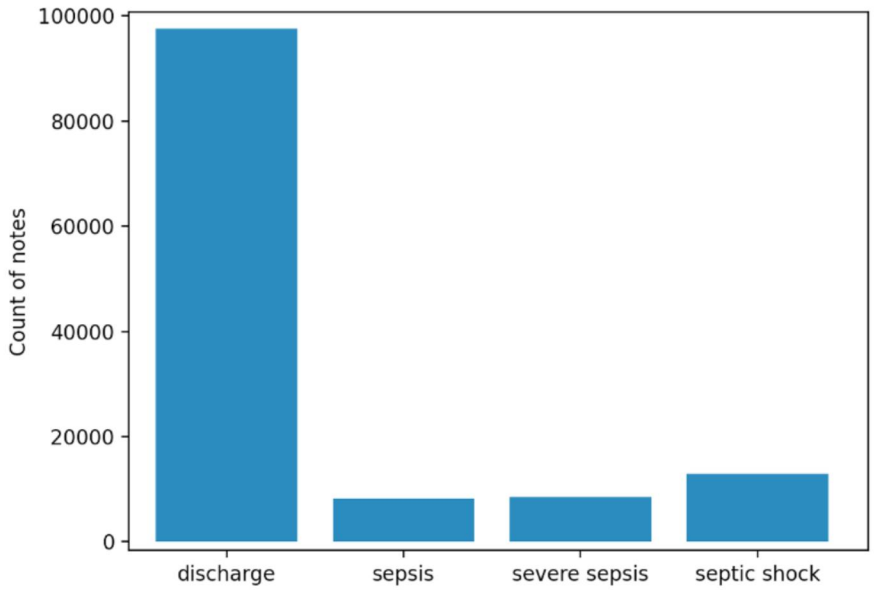
to be the first to apply specific NLP techniques to clinical notes and generate a predictive model. Specifically, the main contribution is in using paragraph vectors to create numeric representations of phrases that allow for semantic comparisons with other phrases within the collection of documents. The paragraph vector methods allows the preservation of word ordering and negations, and so the authors hypothesize that it will perform better than a bag-of-word approach in the prediction model. This paper claims novelty in using these methods on nurse ER triage text in order to develop a prediction model that will determine the eventual needs of the patient, ie hospitalization or discharge.

This study conducted was a retrospective study and this poses a limitation on the analysis conducted. For instance, there is the risk of the outcome in the medical record to be incorrect, and biases within the notes. The vital measurements are currently the standard of care when determining sepsis risk, and so these features must also be included into an automated learning model to be clinically significant.

Additionally, this work draws inspiration from the comparitive study conducted by Henry et al., where they highlight the different methods that are commonly used to predict sepsis [8]. They evaluate a set of methods and find that the highest performing features include physician review, electronic health record billing codes, notes, and presence of comorbidity. We use this idea a the basis for creating a multimodal classification model that uses two of these high performing categories, notes and vital measurements.

3 Dataset and Features

The study uses data from the MIMIC-III dataset, which contains anonymized EHR data from over 50,000 admissions. We first join the patient chart visit with the ICD10 codes to obtain the diagnosis for the each patient stay and we filter out the patients that were diagnosed within the first 3 hours were excluded, because their clinical notes do not fall into the category that is needed for this sepsis study. We consider the earliest time stamped record for each patient, as the goal is to predict sepsis as early as possible. We treat each hospital visit individually. The resulting number of 127126 clinical notes, with a high bias towards notes that correspond to a non-sepsis related diagnosis. There are a total of 29616 sepsis-related notes and 97510 non-sepsis related notes. In order to address this data imabalance, we perform an upsampling procedure that is described within the Methods section. We use BERT [9] to vectorize these sentences, which is also described below in the Methods section.



We then obtain the corresponding vitals measurements for each patient/hospital visit combination that is in our text dataset. We similarly take the corresponding time-stamped vitals data as was used for the text data. We then curated a feature matrix for each patient, which included lab tests, ICD9 codes, vitals, provider notes mapped to SNOMED-CT concepts, number of prescriptions,

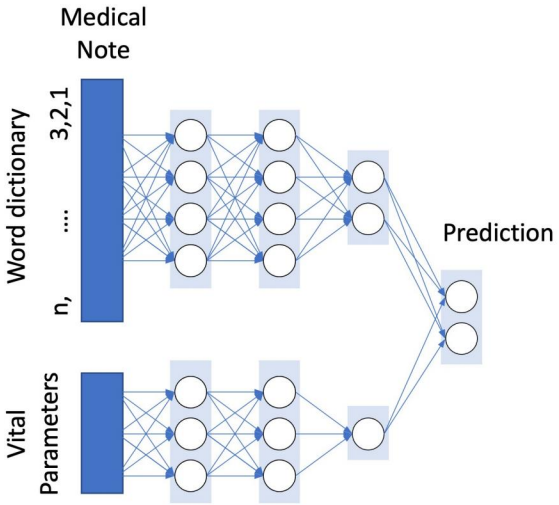
number of within-hospital transfers, subject sex, admission type, and number of abnormal labs. These features are specifically chosen as the entire EHR set is too large to utilize, and these are known to be indicators for developing sepsis.

4 Methods

We begin this study by performing data pre-processing on the dataset. For each note, we first clean the text from stop words, articles, and erroneous punctuation. We then utilize sentenceBert to create a vector for each sentence. Using this method allows for the downstream analysis to utilize the order of words, and the pre-trained embeddings from the repository. In order to combat the class-imabalance of this text classification task, we use an upsampling method to create a uniform distribution of examples within each class. We preturb the examples within the sepsis, severe sepsis, and septic shock cases by choosing each index with probability 0.2 to change its value by 5 percent. In this way we augment the number of sepsis examples within the dataset to equal the number of discharge examples. While we are able to demonstrate positive results using this upsampling method, as shown in the Results section, these types of perturbations are less relevant in the vitals context, as some of the values are discrete options. Therefore, we report the class breakdowns for the text only model, but for further analysis, we combine the sepsis classes into 1 class, allowing for two balanced classes. This still maintains clinical significance, as the ultimate goal is to predict if the patient will be discharged or develop sepsis.

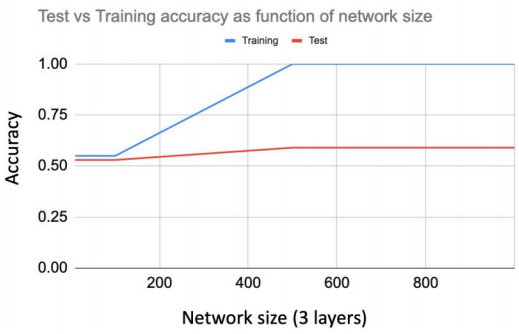
For the vital measurements, we first do the table joins required to get the various datatypes included into the matrix. Since there is missingness in the data, particularly in the lab values, we perform a mean interpolation in order to "fill-in" the missing lab values. As stated earlier, because perturbing the discrete values may introduce significant noise, we instead collapse the three sepsis classes for all the models that involve the vitals data and downsample from the discharge set, leaving two balanced classes.

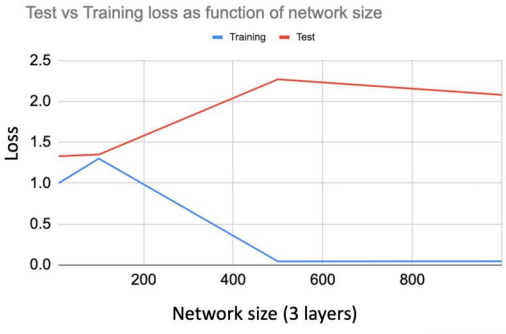
The text vectors are then inputted into a sequential model using 3 deeply connected layers with a reLu activation funtion that end with a softmax layer to predict the 4 classes. A batch size of 128 is used and 150 epochs. The vital measurement matrix is inputted into 3 deeply connected layers with a relu activation function that end with a sigmoid layer to predict the binary classes. A batch size of 128 is used and 150 epochs. In both cases, a cross entropy loss and adam optimizer is used. We tested the vitals model including an attention layer as well, but the test performance decreases, due to overfitting, which we will discuss in the Results section below. Therefore, the model without attention was chosen to be included in the combined network. We also tested various sizes of model, and found varying loss and accuracy values, also discussed in the Results section. The best model was chosen to be implemented as a combined model, using late fusion. In this third model, we use the binary classes, and train the text DNN and vitals DNN together. We again use a categorical-crossentropy loss and adam optimizer. The only difference, is the last layer of text DNN is changed from softmax to sigmoid to perform a binary classification. The model architecture is displayed below.



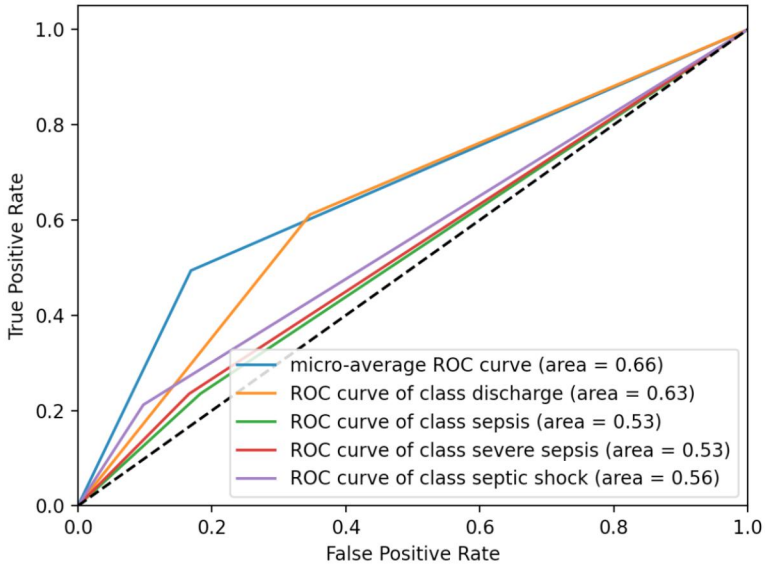
5 Experiments/Results/Discussion

We first experimented with model size to find the optimal number of parameters for each layer, balancing training time and optimal performance. We do this using the balanced text set, training on the DNN architecture described above. We see that the larger sized networks significantly affect the performance of the network, particularly when performing a multi-class prediction task.



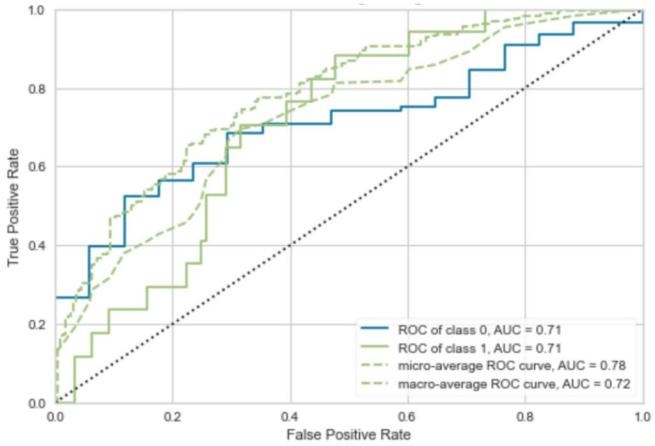


Below, we show the ROC curve for the text only classification model and display the AUC for each class. Here we see that by augmenting the data and making each class balanced, we achieve a more equitable performance across the difference classes, as all show an AUC between 0.5-0.6. However, we can see that this model is not highly performant, probably due to the noise within the data. While cleaning steps were performed on the text data, the text is inherenly noisy and perterbations were added to augment the sepsis classes, leading to a more noisy dataset. Therefore, we can expect that the model does not achieve a high AUC. The training loss was 0.027 and the training accuracy was 0.95, but the test loss was 0.2 and the training accuracy was 0.67. Thus we can see that the model was overfit. Again this can be due to the high variation in text used between examples within the same class and the noise present in the dataset.

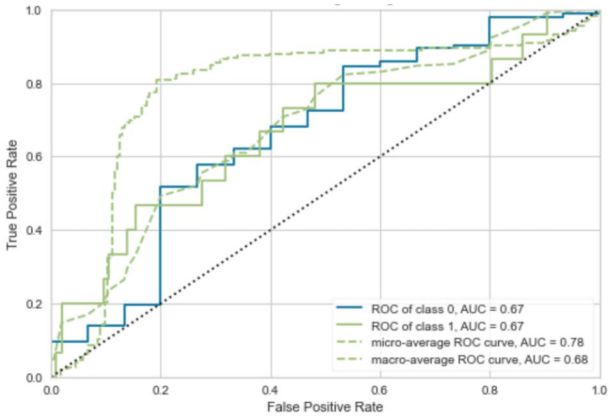


We then experimented with the vitals measurements model architecture, include and without an attention layer. The base model includes a 3-layer DNN, that outputs a binary classification. We modify this to also include an attention layer after the first layer. Without the attention layer, we achieve a test accuracy of 0.87 and with the attention layer, we achieve a test accuracy of 0.76. This is most likely due to the specificity of the features that have already been curated. The features included are those that are biologically and clinically known to impact sepsis. Therefore, applying attention on these curated features causes overfitting of the model. This could be solved with more examples of data points. We utilized dropout layers, but we could also make a larger network with more layers to solve this issue.

For the model without the attention layer, we have shown the ROC and AUC curves for the binary classification task. This model achieves an average UC of 0.72, with an AUC of 0.71 for each class. This shows that the vitals have a better baseline performance than the text. This is due to two reasons. First, the classification problem is simplified when the sepsis classes are merged into one category, and thus the model has a higher performance. Secondly, the vitals dataset is a less noisy dataset as the data present is more structured. Even though we performed a mean interpolation for missing lab values, the effect of this noise is limited when compared to the effect of different writing styles and words chosen when we look at the impact these variations have on the model.



Our final model is a combination of the two previous models, with two 3-layered DNNs and a sigmoid final layer that trains on the vitals and text data together. Despite our expectations With this model, we can see that the performance is decreased compared to the vitals only model, as the average AUC has decreased to 0.68. This is most likely due to the high variance within the text dataset.



6 Conclusion/Future Work

The model based on the vital only data performs the best in this classification task, as it is discrete data that can easily be quantized. To increase the performance of the network with the clinical notes data, we would need to perform more domain-tailored pre-processing steps. While we removed frequently occuring words with little meaning, the ratio between number of words and those that are significant to the diagnosis was still high. So using more domain knowledge to better select the words and sentences to train on. Furthermore, we would like to evaluate the effect of combining the two models at different points in the training algorithm to see the impact, as performing a fusion earlier in the architecture could increase performance.

7 Contributions

Gowri worked on getting the data and Massimo cleaned the data and performed the pre-processing steps. Massimo made the text DNN model and figure and Gowri made the vitals and combined DNN models and figures. We wrote the paper together.

References

References

[1] Peter C Sprivulis, Julie-Ann Da Silva, Ian G Jacobs, George A Jelinek, and Amanda RL Frazer. The association between hospital overcrowding and mortality among patients admitted via western australian emergency departments. *Medical Journal of Australia*, 184(5):208–212, 2006.

[2] Steven Horng, David A Sontag, Yoni Halpern, Yacine Jernite, Nathan I Shapiro, and Larry A Nathanson. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PloS one*, 12(4):e0174708, 2017.

[3] Oleksandr Ivanov, Lisa Wolf, Deena Brecher, Erica Lewis, Kevin Masek, Kyla Montgomery, Yurii Andrieiev, Moss McLaughlin, Stephen Liu, Robert Dunne, et al. Improving ed emergency severity index acuity assignment using machine learning and clinical natural language processing. *Journal of Emergency Nursing*, 47(2):265–278, 2021.

[4] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[5] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. A bert-based universal model for both within-and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71, 2019.

[6] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7(299):299ra122–299ra122, 2015.

[7] Nicholas W. Sterling, Rachel E. Patzer, Mengyu Di, and Justin D. Schrager. Prediction of emergency department patient disposition based on natural language processing of triage notes. *International Journal of Medical Informatics*, 129:184–188, 2019.

[8] Katharine Henry, David Hager, Tiffany Osborn, Albert Wu, and Suchi Saria. Comparison of automated sepsis identification methods and electronic health record–based sepsis phenotyping: Improving case identification accuracy by accounting for confounding comorbid conditions. *Crit Care Explor.*, 2019.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.