

Project Proposal: Predicting Hospital Admissions Using MIMIC-IV EHR Data with BERT

Yufei Deng (yd345)

Yihan Liu (yl2548)

Hang Shi (hs952)

March 8, 2024

1 Introduction

Emergency Departments (ED) are critical components of healthcare systems, serving as the first point of contact for patients with a wide range of conditions. However, the efficiency of EDs is often hampered by the time and effort clinicians must invest in assessing and prioritizing patient cases, especially non-urgent ones. The challenge lies in quickly identifying the urgency of a patient's condition to allocate resources effectively and improve patient outcomes. This project proposes the development of an NLP-based system to automatically classify the urgency of patients coming to the ED, using the data available in electronic health records (EHRs).

2 Related Work

Recent advancements in NLP have shown promising results in healthcare, particularly in processing and interpreting unstructured clinical text. A notable contribution to this field is the work by Rasmy et al. (2021), which demonstrates the effectiveness of BERT models in understanding complex medical narratives for disease prediction. Inspired by their findings, this project extends the application of BERT models to the domain of emergency care, focusing on the task of urgency grading of ED admissions. While Rasmy et al. laid the groundwork for using BERT in clinical contexts, our project aims to leverage these insights specifically for the prioritization of patient care in emergency settings.

3 Approach

The proposed project will implement a Bidirectional Encoder Representations from Transformers (BERT) model to classify patients based on the urgency of their conditions. The BERT model will be fine-tuned on the dataset extracted from MIMIC-IV, which includes comprehensive admission records from the ED. These records encompass patient information, chief complaints, medical history, diagnosis codes, and radiology reports, providing a multifaceted view of each patient's case.

To prepare the data for model training, we will preprocess the text to extract key features such as patient demographics, medications, and severity of the complaint in the clinical documentation. The BERT model will then be trained to predict whether a patient will be admitted to the hospital based on their EHR data.

3.1 Experimental Plan

To validate the effectiveness of the BERT-based approach, we will compare its performance against several baseline models, including:

- Multiclass Logistic Regression: We will use logistic regression with features extracted from the admission dataset. Feature extraction methods will include basic NLP techniques and sentiment classification to gauge the severity implied in the textual data.
- Other Baseline Models: Depending on initial findings, we may explore additional models such as SVM or simple RNNs for comparative analysis.

The dataset will be divided into training, validation, and testing sets as provided in the MIMIC-IV split (109,168 admissions). Performance metrics will include accuracy, precision, recall, and F1 score, with a particular focus on the model's ability to correctly identify high-urgency cases.

3.2 Expected Outcomes and Impact

We will build a classification model using BERT that can accurately predict ED admissions, potentially reducing the time clinicians spend on non-urgent cases. This tool could significantly improve ED operations by optimizing patient flow and ensuring that critical cases receive immediate attention. Furthermore, the insights gained from this project could contribute to the broader application of NLP techniques in healthcare, especially in processing and extracting actionable insights from unstructured clinical data.

References

- [1] Rasmy, L., Xiang, Y., Xie, Z. et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.* 4, 86 (2021). <https://doi.org/10.1038/s41746-021-00455-y>
- [2] Johnson, Alistair, et al. "MIMIC-IV" (version 0.4). *PhysioNet* (2020), <https://doi.org/10.13026/a3wn-hq05>.