Preserving the Knowledge of Long Clinical Texts Using Aggregated Ensembles of Large Language Models

Mohammad Junayed Hasan[1], Suhra Noor[1], Mohammad Ashrafuzzaman Khan[1]
[1]Department of Electrical and Computer Engineering, North South University, Bashundhara, Dhaka

2 Nov 2023

Abstract

Clinical texts, such as admission notes, discharge summaries, and progress notes, contain rich and valuable information that can be used for various clinical outcome prediction tasks. However, applying large language models, such as BERT-based models, to clinical texts poses two major challenges: the limitation of input length and the diversity of data sources. This paper proposes a novel method to preserve the knowledge of long clinical texts using aggregated ensembles of large language models. Unlike previous studies which use model ensembling or text aggregation methods separately, we combine ensemble learning with text aggregation and train multiple large language models on two clinical outcome tasks: mortality prediction and length of stay prediction. We show that our method can achieve better results than baselines, ensembling, and aggregation individually, and can improve the performance of large language models while handling long inputs and diverse datasets. We conduct extensive experiments on the admission notes from the MIMIC-III clinical database by combining multiple outputs and high-dimensional datasets, demonstrating our method's effectiveness and superiority over existing approaches. We also provide a comprehensive analysis and discussion of our results, highlighting our method's applications and limitations for future research in the domain of clinical healthcare. The results and analysis of this study is supportive of our method assisting in clinical healthcare systems by enabling clinical decision-making with robust performance overcoming the challenges of long text inputs and varied datasets.

Keywords: Clinical outcome prediction, Large language models, Text aggregation, Ensemble learning, Clinical texts

1 Introduction

Clinical Natural Language Processing (NLP) has revolutionized the healthcare industry by providing effective ways to interpret and analyze the vast amount of clinical data [1]. Usually clinical data are stored in the form of Electronic Health Records (EHRs) in massive databases like the MIMIC-III clinical database [2] or eICU clinical database [?], which are usually unstructured, high-dimensional, heterogeneous, temporally dependent, irregular and contain sparse information [?]. Clinical texts, specifically patients' admission notes, are rich in critical information that can inform healthcare decisions and improve patient outcomes. These notes can prevent doctors from overlooking possible risks and help hospitals to plan capacities. Moreover, they can also be used in clinical diagnosis, treatment, decision support, information retrieval, and knowledge discovery.

...

References

[1] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare: management, analysis and future prospects. *Journal of big data*, 6(1):1–25, 2019.

[2] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

...