Review of: "NER Sequence Embedding of Unified Medical Corpora to Incorporate Semantic Intelligence in Big Data Healthcare Diagnostics"

Hanane Allioui

Potential competing interests: No potential competing interests to declare.

This paper discusses the challenges of clinical diagnosis and the importance of integrating technology into the process. It emphasizes the need for a semantic understanding of the medical domain and clinical context to enable intelligent analytics and improve diagnostic accuracy. The paper also highlights the prevalence of diabetes mellitus (DM) and the lack of unified corpora tagged with medical terms for training analytics in diagnosing DM and its comorbidities. The research aims to address this gap by collecting endocrine diagnostic electronic health records (EHR) corpora labeled with the ICD-10-CM coding scheme. Different NER sequence embedding approaches using machine learning and natural language processing techniques were experimented with, resulting in a maximum accuracy of 0.9 for a specific corpus. The paper emphasizes the significance of clinical notes and practitioner comments in improving diagnostic accuracy. It seems that the paper presents a solid and interesting method. The results show that the proposed process provides promising results. However, I believe that some minor issues need to be addressed before publication. Please find them below:

1. The text is clear and straightforward in its presentation. It effectively communicates the challenges of clinical diagnosis, the importance of technology integration, and the research's motivation and objectives. The text is well-structured, with a logical flow from identifying the problem to discussing the research approach and findings.
2. The paper correctly identifies the challenge in the medical domain, particularly the lack of unified corpora for training analytics in diagnosing DM and comorbidities, which is a significant issue in healthcare. However, it would be helpful to state the overall purpose or contributions of the research explicitly.
3. The methodology is well described. Yet, expanding the description of the NER (Named Entity Recognition) sequence embedding approaches and machine learning techniques used in the research would be valuable. Explain clearly why these specific methods were chosen and how they were applied to the data.
4. Results and Significance: The paper reports a maximum accuracy of 0.9 for a specific corpus, which is a promising result. The emphasis on the importance of clinical notes and practitioner comments in enhancing diagnostic accuracy adds value to the research. However, authors should discuss how they intend to enhance the accuracy.
5. The study needs more information about the practical implications of the research. Including a brief discussion on that, would add functional value to the investigation.
6. I have found some grammatical errors, kindly proofread the manuscript.
7. High image quality is essential, please provide better quality.

8. As the paper presents an important process, it is required that the discussion section should be presented in a clear and separate section.