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

Zina Ben Miled[1]

[1] Indiana University/Purdue University at Indianapolis

Potential competing interests: No potential competing interests to declare.

The application focus of the manuscript is important and challenging and the authors are cognizant of the current litterature in the field. However, the manuscript can benefit from a better structure and more focus. In its current version, it reads like a combination of multiple papers. If the objective is to develop a multi-class classifier using a customized embedding then the authors should focus on this aspect.

The related work section should omit some of the well understood concepts and focus more on recent development in the field. More recent embedding language models should be considered as mentioned in the Future Work section by the authors. LMs such as ClinicalBERT, BioBERT and MedBert should be considered. The same is true for other sections, not enough details is included for important sections (e.g., sections 3.3, 3.4, 3.5) and less important details are included in other sections (e.g., Section 4.3)

There several statements in the manuscript that were not justified. For example “In our study, we omitted the use of POS tagging and applied NER embedding only that minimized the code and kept it simple.”

Were the number of comorbities specifically truncated? For example, the 100 patients 2K records has only 5 comorbidities but the 100 patients 9K records has 65 comorbidities whereas the 14K patient 33K recods has 30 comorbidities. Varying the size of the cohort while keeping the number of classes fixed and vice-a-versa can help with the interpretation of the results.

Statements such as "This way of preprocessing clinical notes and practitioner comments did not let sentence length and vocabulary affect the diagnostic results and performance of the ML models." are not well explained.

I am not sure that the comparison of the different cloud platform is suitable for this manuscript.

The caption and format of figures and tables can be improved. The manuscript can also benefit from English editing.