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

Youlin Zhao[1,2]

1 Hohai University, Nanjing, China
2 Nanjing University, China

Potential competing interests: No potential competing interests to declare.

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

Accept after major revision and modification

This paper focuses on the problem of automated clinical diagnosis, especially in the identification of diabetes and its comorbidities, which has significant application value in the medical field, but there are some problems as follows

1 Introduction

1. The introduction is too long; multiple sentences are too informative and loaded with too many technical terms, which reduces readability.
2. Do not specify how to evaluate the performance of the proposed model.

2 Related Work

1. The content is piled up too much, lacking a clear logical level. For example, the historical development of NLP tools, the classification of embedded models, and the use of medical corpora can be discussed in detail as subsections, respectively, to improve the readability of chapters.
2. This chapter is mainly a summary of the existing work and lacks an in-depth discussion of its shortcomings or room for improvement. For example, although different embedding models are mentioned, their performance differences in a medical context are not evaluated in detail.
3. The introduction of some research content is abrupt, and its relationship to the overall research objective is not clearly stated. At the beginning and end of each section, it is recommended to add a description of the relevance to the overall research topic. For example, can you explain how these tools or techniques provide the basis for the research in this article?

3 Architecture and Design

1. While anonymization is mentioned in Section 3.1, you could elaborate on the specific methods used to ensure patient

data security and privacy, especially given the sensitivity of healthcare data.

2. The results mentioned the influence of different corpus sizes on the accuracy of the model, but the explanation of this phenomenon was not deep enough. It is recommended that the authors further discuss the specific effects of corpus size on model performance, especially in the case of missing data or missing records, how to deal with these data issues, and their potential impact on results.
3. It is recommended to provide a more specific analysis of the effects of oversampling and undersampling techniques to explain why their effects are more limited.

6 TensorFlow.Keras NER Embeddings using proposed Bi-LSTM Dense Layered Neural Networks Approach

1. The paper mentions the choice of hyperparameters (e.g., learning rate, optimizer, activation function) but lacks a detailed justification for these choices. Including a more thorough explanation of how these hyperparameters were selected, possibly through grid search or cross-validation, would add more credibility to the model's design.

8 Conclusion and Future Work

1. Although the paper mentions the possibility of expanding the unified knowledge base and automating ICD-10 coding, it would be useful to delve deeper into future challenges and broader applications of the proposed method. How might the framework scale for more complex datasets or other diseases? What specific steps are needed to make this approach more robust for real-world clinical use?