REPORT

Alvi Anowar
ID: 23366031

Paper Title: DiLBERT: Cheap Embeddings for Disease-Related Medical NLP

1. Summary

1.1 Motivation

The manual coding process in the biomedical field is time-consuming and error-prone, leading to a growing demand for clinical text-mining tools. Although Natural Language Processing (NLP) models have shown effectiveness at the sentence level, there is a recognized necessity for specialized models in the biomedical domain, especially for handling diagnostic expressions.

1.2 Contribution

This paper presents a specialized BERT model for diagnoses and health conditions tasks. The authors created a pre-training corpus using entities from the ICD-11 classification to leverage the transfer learning capabilities of NLP models.

1.3 Methodology

The methodology involves the creation of a disease-related language model through pre-training on a corpus derived from ICD-11 entities and supplemented with relevant documents from PubMed and Wikipedia. Fine-tuning was subsequently performed for three downstream tasks using two different datasets. Despite being trained on a smaller corpus than state-of-the-art algorithms, the model demonstrated comparable or higher accuracy scores across all functions.

1.4 Conclusion

The model achieved a 97.53% accuracy on death certificate coding and an 81.32% accuracy on clinical document coding, slightly surpassing other models. Despite the model being trained on a smaller corpus, its performance equals or surpasses that of existing state-of-the-art models.

2. Limitation

2.1 First Limitation

The presented work is focused on tasks related to ICD-11 entities and the lack of testing on functions associated with classifications or terminologies other than ICD.

2.2 Second Limitation

The current work adheres to a one-stage fine-tuning process and may not capture the full potential of more intricate fine-tuning strategies.

3. Synthesis

The experimental results highlight the superiority of the proposed domain-specific BERT model over existing models in similar tasks. Its strong performance, coupled with minimal training

requirements, suggests practicality for developing models in languages beyond English. Future research should explore multi-stage fine-tuning processes for the domain-specific BERT model, understanding their effects and potential complications. Additionally, further work is recommended to test the model on tasks related to different terminologies as datasets become available.