Title: DiLBERT: Cheap Embeddings for Disease Related Medical NLP

Authors: Kevin Roitero, Beatrice Portelli, Mihai Horia Popescu, and Vincenzo Della Mea

Institution: Department of Mathematics, Computer Science, and Physics, University of Udine, Italy

Abstract: The specialised Bidirectional Encoder Representations from Transformers (BERT) model DiLBERT is introduced in this study. It is intended for use in medical Natural Language Processing (NLP) with a focus on problems connected to diseases. DiLBERT, in contrast to general-purpose models, uses a narrower, disease-related corpus for pre-training and shows impressive performance on tasks and embedding generation efficiency. The study addresses important issues including resource limitations and domain-specific subtleties in medical NLP, demonstrating the applicability of DiLBERT in accurately classifying medical texts into suitable classifications like the International Classification of Diseases (ICD-11).

Introduction: A vital component of healthcare administration, electronic health records (EHRs) hold a plethora of information about diagnosis and medical disorders. The study recognises the increasing need for effective and precise clinical text mining technologies, especially for standardising the categorization of medical texts. In answer to the demand for more specialised language models in the medical field, the researchers present DiLBERT.

Methodology: Using a disease-related corpus that is derived from ICD-11 entities and enhanced with content from PubMed and Wikipedia, DiLBERT is pre-trained. Using two separate datasets, the model is fine-tuned for three downstream tasks. According to the study, this targeted strategy makes better use of available data and resources, which could make model creation easier even for languages other than English because fewer data are required.

Results: The outcomes demonstrate the efficacy and efficiency of DiLBERT. It obtains excellent accuracy scores on all activities examined, with a focus on clinical document coding and death certificate coding. Despite being trained on a significantly smaller corpus, the model performs on par with or better than state-of-the-art methods. These findings highlight how DiLBERT may transform medical NLP by giving academics and healthcare providers precise, practical, and easily accessible tools.

Even though DiLBERT represents a substantial achievement, the study addresses the problems that still need to be solved as well as possible future possibilities. The researchers noted the model's reliance on the calibre and size of the training corpus and the requirement for thorough validation and testing across a range of languages and subjects.

Conclusion: A strong argument is made for the effectiveness and promise of specialised language models in medical informatics in "DiLBERT: Cheap Embeddings for Disease Related

Medical NLP". The study lays the groundwork for future research and development in the field of medical natural language processing (NLP) by demonstrating how a focused, domain-specific strategy can result in considerable gains in task performance and efficiency. DiLBERT is proof of the inventive spirit and unwavering quest for improved healthcare via technology.