DEEP LEARNING EMBEDDING MODELS FOR MULTIMORBIDITY DETECTION USING HEALTHCARE ADMINISTRATIVE DATABASES

Koumantakis Emanuele[1,2,3], Visconti Alessia[2], Emanuele Pietropaolo[2], Sandro Gepiro Contaldo[4], Marco Beccuti[4], Berchialla Paola[2], TrustAlert Group

[1] Department of Public Health and Pediatrics, University of Torino, 10126 Torino, Italy
[2] Centre for Biostatistics, Epidemiology and Public Health, Department of Clinical and Biological Sciences, University of Torino, 10043 Orbassano, Italy
[3] Post Graduate School of Medical Statistics, University of Torino, 10126 Torino, Italy
[4] Department of Computer Science, University of Torino, 10124 Torino, Italy

Introduction

Traditional machine learning and statistical models frequently face challenges when dealing with high-dimensional data and irregular time intervals between events, both of which are common features of Healthcare Administrative Databases (HADs). Deep Learning (DL) techniques, particularly those involving natural language processing (NLP), are well-suited to manage the extensive data found in HADs. These models can learn concise representations of diagnosis sequences over time with minimal pre-processing, making them powerful tools for analysing and interpreting healthcare data.

Aim

The goal of this study is to train a NLP model on structured HADs to generate embeddings for detecting multimorbidity.

Methods

The study included patients aged 65 or older from the ASL-CN 2, with longitudinal observations including diagnosis and medication sequences. This age group was selected due to the higher prevalence of multimorbidity among older adults. To capture the inherent structure of these observations, the original Google Bidirectional Encoder Representations from Transformers (BERT) model was pre-trained for two self-supervised tasks: the Masked Language Model (MLM) and the Next Sentence Prediction (NSP) task [1]. In the MLM task, words within the sequences were randomly masked, and BERT was trained to predict the original words from their surrounding context, using information from both the left and right words. The NSP task involved predicting whether one sentence (or patient visit) logically follows another, which mirrors the sequential nature of patient visits and medical events over time. The quality of the generated embeddings was assessed using a t-SNE (t-distributed Stochastic Neighbor Embedding) analysis, a technique that visualizes high-dimensional data by reducing it to two (or three) dimensions [2].

Results

The pre-trained BERT model demonstrated high accuracy, achieving 96.58% (SE = 0.02) for the MLM task and 99.75% (SE = 0.005) for the NSP task. These results indicate that BERT effectively learned the structure of HADs' data, capturing the complex relationships between different medical events and treatments. The embeddings enabled the identification of co-occurring pathologies and polypharmacy conditions. Notable patterns of multimurbidity identified include: (i) rehabilitation, osteoarthritis, and connective tissue diseases, (ii) respiratory failure with respiratory and dysrhythmia diagnosis, (iii) anti-inflammatory/anti-rheumatic and anti-thrombotic drugs with reflux medications, and (iv) beta-lactam/penicillin antibiotics, anti-inflammatory/anti-rheumatic, quinolone antibacterial drugs and other bacterial infections diagnoses.

Conclusions

The high accuracy of the pre-trained model indicates that BERT can efficiently learn the complex data structures found in HADs. Training on the MLM and NSP tasks achieved very satisfactory performance, with at least 97% accuracy. The t-SNE analysis revealed that co-occurring medical prescriptions and hospitalization diagnoses have similar medical meanings (e.g., thyroid preparations are linked to thyroid disorders). Overall, this study demonstrates the potential of using BERT and other NLP models to analyze structured healthcare data, providing a powerful tool for identifying patterns of multimorbidity and improving patient care. Future research could explore the application of these techniques to other types of healthcare data and investigate their potential for prognosis detection and personalized medicine.

References

1. Abadi M., Agarwal A., Barham P., et al.,.TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems [Internet]; 2015; http://tensorflow.org/
2. van der Maaten L.J.P., Hinton G.E., Visualizing High-Dimensional Data Using t-SNE; Journal of Machine Learning Research, 2008; 9(86):2579–2605.

Table 1. Performance of the model in terms of accuracy in pre-train tasks (8 epochs further pre-train). Abbreviations: BERT, Bidirectional Encoder Representations from Transformers; HAD, Healthcare Administrative Databases; MLM, Masked Language Modeling; NSP, Next Sentence Prediction; F1, Harmonic mean of precision and recall.

Metric		**BERT model**		
pre-trained		✗	✓	✓
HAD pre-train		✓	✗	✓
MLM & NSP	Accuracy	70.60%	68.12%	**86.76%**
	F1	82.09%	81.04%	**89.84%**
MLM only	Accuracy	67.60%	68.55%	**87.02%**
	F1	80.67%	81.34%	**91.04%**
NSP only	Accuracy	69.12%	**87.72%**	68.99%
	F1	81.74%	**90.99%**	81.65%