

ICD (*International Classification of Diseases*) versi kesepuluh adalah format standar pemberian nama diagnosis pada *Electronic Medical Record* (EMR) yang digunakan di Indonesia. Pelabelan ICD-10 dilakukan secara *multilabel* karena seorang pasien dapat tercatat memiliki lebih dari satu kode ICD-10, yang menunjukkan adanya beberapa diagnosis penyakit. Pelabelan ICD-10 ini dilakukan dengan dua metode. Dokter dapat menuliskan langsung kode ICD-10 dengan panduan kamus ICD-10, atau dokter mencatat informasi medis yang kemudian dimanfaatkan oleh petugas koding. Proses ini menghadapi tantangan terkait kompleksitas struktur EMR, jumlah kode yang banyak, dan kemiripan makna antar kode. Hal ini semakin rumit pada kasus-kasus mental, di mana informasi sering kali berupa narasi panjang. Oleh karena itu, diperlukan sebuah teknologi yang dapat memberikan rekomendasi pengambilan keputusan pelabelan ICD-10. Penelitian ini bertujuan untuk mengimplementasikan teknologi *transformer* berupa BERT dalam klasifikasi *multilabel* ICD-10 berdasarkan catatan medis dari EMR. *Dataset* EMR yang digunakan berasal dari RSUP Dr Sardjito, mencakup rawat inap dan rawat jalan di instalasi poli serta bangsal penyakit jiwa. Penelitian memanfaatkan model *pretrained* BERT berbahasa Indonesia, yaitu IndoBERT, mBERT, dan DistilBERT. *Grid search* dan *hyperparameter tuning* digunakan untuk menemukan konfigurasi terbaik dari setiap model. Hasil penelitian menunjukkan bahwa mBERT memberikan performa terbaik dengan F1-score makro sebesar 35,35%. Optimalisasi model melalui *hyperparameter tuning* dan *grid search* terbukti efektif dalam meningkatkan performa. Berdasarkan hasil statistik, penerapan mekanisme *label aware attention* pada DistilBERT dan IndoBERT tidak memberikan pengaruh signifikan terhadap hasil klasifikasi. Selain itu, penerapan *label aware attention* menurunkan performa mBERT.

Kata kunci: *natural language processing*, *multilabel*, BERT, EMR, ICD-10

ABSTRACT

ICD (International Classification of Diseases) tenth version is a standard format for naming diagnoses in Electronic Medical Record used in Indonesia. ICD-10 labeling is conducted in a multilabel manner because a patient may have more than one ICD-10 code recorded, indicating multiple diagnoses. ICD-10 labeling is carried out through two methods. Physicians can directly write down the ICD-10 code using the ICD-10 dictionary guide, or they can record medical information, which is then utilized by coding staff. This process faces challenges related to the complexity of the EMR structure, the large number of codes, and the similarity in code meanings. It becomes even more complicated in mental health cases, where information is often in the form of lengthy narratives. Therefore, a technology that can provide decision-making recommendations for ICD-10 labeling is required. This study aims to implement transformer technology for multilabel classification of ICD-10 based on medical record features or information found in EMR. The EMR dataset used originates from Hospital X, covering inpatient and outpatient categories in the polyclinic and psychiatric ward. The study utilizes pretrained BERT models in Indonesian, namely IndoBERT, mBERT, and DistilBERT. Grid search and hyperparameter tuning are applied to determine the optimal configuration for each model. The results indicate that mBERT provides the best performance, with a macro F1-score of 35.35%. Model optimization through hyperparameter tuning and grid search proves effective in improving performance. Based on statistical results, the application of the label aware attention mechanism on DistilBERT and IndoBERT does not significantly impact classification results, and in fact, decreases the performance of mBERT.

Keywords : natural language processing, multilabel, BERT, EMR, ICD-10