INTISARI

Perbandingan Performa Pre-trained BERT untuk Named Entity Recognition pada Dataset Uji Kelayakan Klinis

Oleh

Muhammad Rayi

20/455450/PA/19665

Perkembangan teknologi dan informasi, salah satunya pada sektor klinis menghasilkan kemampuan yang disebut interoperabilitas. Interoperabilitas merupakan kemampuan 2 sistem/lebih untuk melakukan pertukaran data. Namun, masalah muncul pada tipe data tidak terstruktur, seperti catatan klinis yang menghambat proses interoperabilitas. Terkait hal ini, maka dibutuhkan model yang berperforma tinggi dalam memproses tipe data ini untuk tugas *named entity recognition*, yaitu menggunakan 3 model *pre-trained* BERT domain klinis, yaitu PubMedBERT, ClinicalBERT, dan MedBERT.

Proses penelitian dilakukan dengan melakukan *fine-tuning* antara ketiga model dengan dataset Chia, yang merupakan dataset uji kelayakan klinis yang berupa anotasi dan *free text.* Proses *fine-tuning* dilakukan dengan *framework* Optuna untuk mencari kandidat *hyperparameter* terbaik untuk setiap model. Selain itu, dilakukan modifikasi fungsi *loss* untuk mengatasi permasalahan *imbalanced* pada dataset Chia.

Hasil penelitian menunjukkan bahwa model PubMedBERT dan ClinicalBERT mempunyai hasil *f1 score* yang hampir setara. PubMedBERT mengungguli *f1 score* pada 8 label dibanding kedua model lainnya sedangkan ClinicalBERT mengungguli secara *macro average f1 score* sebesar 64%. Selain itu, didapatkan waktu *testing* tercepat oleh PubMedBERT serta memori terkecil oleh ClinicalBERT dan MedBERT. Hasil eksperimen juga menunjukkan perubahan fungsi *loss* dapat meningkatkan *f1 score* pada label minoritas.

Kata Kunci: BERT, interoperabilitas, *named entity recognition*, catatan klinis, data terstruktur, dan data tidak terstruktur.

ABSTRACT

COMPARISON OF PRE-TRAINED BERT PERFORMANCES FOR NAMED ENTITY RECOGNITION ON CLINICAL TRIAL ELIGIBILITY DATASET

by

Muhammad Rayi

20/455450/PA/19665

The development of technology and information, particularly in the clinical sector, has led to a capability known as interoperability—the ability of two or more systems to exchange data. However, challenges arise with unstructured data such as clinical notes, which hinder this process. To address this, high-performance models are needed to process such data for named entity recognition, using three pre-trained clinical-domain BERT models: PubMedBERT, ClinicalBERT, and MedBERT.

The study fine-tuned these three models on the Chia clinical trial eligibility dataset, which consists of annotations and free text. Fine-tuning was conducted using the Optuna framework to search for the best hyperparameters for each model, along with a modified loss function to address class imbalance.

Results show that PubMedBERT and ClinicalBERT achieved almost equal F1 scores. PubMedBERT outperformed on 8 labels and had the fastest testing time, while ClinicalBERT achieved the highest macro-average F1 score (64%) and, along with MedBERT, the lowest memory usage. The modified loss function also improved F1 scores for minority labels.

Keywords: BERT, interoperability, named entity recognition, clinical notes, structured data, and unstructured data.