

INTISARI

PENINGKATAN KEMAMPUAN *ZERO-SHOT* GLiNER UNTUK NER MEDIS BAHASA INDONESIA MELALUI AUGMENTASI DATA SINTETIS BERBASIS LLM DAN *FINE-TUNING*

Oleh

Aisyah Putri Khurin'in

22/493908/PA/21227

Named Entity Recognition (NER) pada teks konsultasi kesehatan daring berbahasa Indonesia menghadapi tantangan berupa keterbatasan data berlabel, variasi bahasa informal, dan kompleksitas istilah medis. Penelitian ini meningkatkan kinerja *Generalist and Lightweight Model for Named Entity Recognition* (GLiNER) melalui augmentasi data sintetis berbasis *Large Language Models* (LLM) dan *fine-tuning*. Data sintetis dihasilkan menggunakan Gemini 2.5 Flash dengan *prompt* terstruktur sesuai skema anotasi dataset telemedisin Indonesia dan dievaluasi pada dataset `nadhifikbarw/id_ner_nimas` menggunakan *precision*, *recall*, dan *F1-score*. Hasil menunjukkan bahwa seluruh skenario *fine-tuning* meningkatkan performa secara konsisten dibandingkan *baseline zero-shot* (micro $F1=0,266$). Performa terbaik dicapai pada *full fine-tuning* dengan dataset asli (micro $F1=0,768$), sedangkan *fine-tuning* dengan gabungan data asli dan sintetis mencapai micro $F1=0,622$. Penelitian ini membuktikan bahwa data sintetis berbasis LLM efektif sebagai alternatif data latih pada kondisi *low-resource* dan memperkuat potensi GLiNER untuk ekstraksi informasi medis dalam layanan telemedisin Indonesia.

Kata Kunci:

Named Entity Recognition (NER), GLiNER, Data Sintetis, *Large Language Models* (LLM), *Zero-Shot Learning*, *Fine-Tuning*, Bahasa Indonesia, Domain Medis, Telemedisin

ABSTRACT

ENHANCING ZERO-SHOT GLiNER CAPABILITY FOR INDONESIAN MEDICAL NAMED ENTITY RECOGNITION (NER) THROUGH LLM- BASED SYNTHETIC DATA AUGMENTATION AND FINE-TUNING

Oleh

Aisyah Putri Khurin'in

22/493908/PA/21227

Named Entity Recognition (NER) applied to Indonesian online health consultation texts faces challenges including limited labeled data, informal language variation, and complex medical terminology. This study improves the performance of the Generalist and Lightweight Model for Named Entity Recognition (GLiNER) through LLM-based synthetic data augmentation and fine-tuning. Synthetic data were generated using Gemini 2.5 Flash with structured *prompts* following the annotation scheme of the nadhifikbarw/id_ner_nimas telemedicine dataset, evaluated using precision, recall, and F1 score. All fine-tuning scenarios consistently outperformed the zero-shot baseline (micro F1=0.266). The best result was achieved by full fine-tuning on original data (micro F1=0.768), while fine-tuning on a combination of original and synthetic data reached micro F1=0.622. These results confirm that LLM-generated synthetic data is a viable training alternative in low-resource conditions and support the potential of GLiNER for medical information extraction in Indonesian telemedicine services.

Keywords:

Named Entity Recognition (NER), GLiNER, Synthetic Data, Large Language Models (LLM), Zero-Shot Learning, Fine-Tuning, Indonesian Language, Medical Domain, Telemedicine