

INTISARI

ANALISIS *TRADE-OFF* DALAM PENERAPAN *KNOWLEDGE DISTILLATION* PADA MODEL *QUERYNER* DENGAN INTEGRASI *CONDITIONAL RANDOM FIELDS* UNTUK *NAMED ENTITY RECOGNITION* BAHASA INGGRIS

Oleh

Muhammad Dafa Wisnu Galih

22/503607/PA/21649

Dengan meningkatnya volume transaksi pada *platform e-commerce*, kebutuhan akan sistem pencarian produk yang efisien dan responsif menjadi semakin krusial. Model berbasis Transformer seperti QueryNER telah menunjukkan performa yang tinggi dalam tugas *Named Entity Recognition* (NER), tetapi memiliki keterbatasan berupa kompleksitas komputasi yang besar dan waktu inferensi yang tinggi. Penelitian ini bertujuan untuk meningkatkan efisiensi model NER melalui penerapan metode *Knowledge Distillation* (KD) berbasis respons, dengan mentransfer pengetahuan dari model *teacher* QueryNER ke model *student* yang lebih ringan, yaitu DistilBERT, TinyBERT, dan BiLSTM. Selain itu, penelitian ini juga mengeksplorasi pengaruh penambahan lapisan *Conditional Random Field* (CRF) terhadap konsistensi label dan performa model. Eksperimen dilakukan melalui pelatihan model *baseline*, penerapan KD dengan berbagai konfigurasi parameter, serta evaluasi menggunakan metrik berbasis token dan entitas. Analisis efisiensi dilakukan dengan mengukur latensi inferensi, *throughput*, jumlah parameter, dan FLOPs. Hasil penelitian menunjukkan bahwa KD efektif dalam mempertahankan performa model berbasis Transformer dengan kompleksitas yang lebih rendah. Model DistilBERT tanpa CRF hasil KD memberikan performa terbaik pada data uji sekaligus memiliki latensi yang jauh lebih rendah dibandingkan model *teacher*. Sebaliknya, KD kurang memberikan peningkatan pada model BiLSTM. Penambahan CRF tidak memberikan peningkatan signifikan pada model Transformer, tetapi mampu meningkatkan konsistensi prediksi pada model BiLSTM dalam evaluasi berbasis entitas, meskipun dengan konsekuensi peningkatan latensi yang signifikan. Penelitian ini menunjukkan bahwa kombinasi KD dan model Transformer ringan merupakan pendekatan yang efektif untuk menghasilkan model NER yang efisien dan tetap kompetitif untuk aplikasi *e-commerce*.

Kata-kata kunci: *Named Entity Recognition, Knowledge Distillation, QueryNER, Transformer, Conditional Random Field, e-commerce.*

ABSTRACT

TRADE-OFF ANALYSIS IN APPLICATION OF KNOWLEDGE DISTILLATION ON QUERYNER MODEL WITH CONDITIONAL RANDOM FIELDS INTEGRATION FOR ENGLISH NAMED ENTITY RECOGNITION

By

Muhammad Dafa Wisnu Galih

22/503607/PA/21649

With the increasing volume of transactions in e-commerce platforms, the need for efficient and responsive product search systems has become increasingly critical. Transformer-based models such as QueryNER have demonstrated strong performance in Named Entity Recognition (NER) tasks; however, they suffer from high computational complexity and latency during inference. This study aims to improve model efficiency by applying response-based Knowledge Distillation (KD), transferring knowledge from a large and complex teacher model (QueryNER) to smaller and more efficient student models, namely DistilBERT, TinyBERT, and BiLSTM. Additionally, this study investigates the impact of incorporating a Conditional Random Field (CRF) layer to enhance label consistency and prediction performance. Experiments were conducted through baseline model training, KD implementation with various parameter configurations, and evaluation using both token-level and entity-level metrics. Model efficiency was assessed based on inference latency, throughput, parameter size, and FLOPs. The results show that KD effectively preserves the performance of Transformer-based student models while significantly reducing computational cost. The best configuration is achieved by DistilBERT without CRF, which outperforms the teacher model on test data while maintaining substantially lower latency. In contrast, KD provides limited improvement for the BiLSTM model. The addition of CRF does not significantly improve Transformer-based models but slightly enhances entity-level consistency in BiLSTM, albeit with a considerable increase in inference latency. These findings demonstrate that combining knowledge distillation with lightweight Transformer models is an effective approach for developing efficient and practical NER systems in the e-commerce domain.

Keywords: Named Entity Recognition, Knowledge Distillation, QueryNER, Transformer, Conditional Random Field, e-commerce.