UNIVERSITAS
GADJAH MADA
Pengembangan Model Passage Reranking Berbasis BERT Cross Encoder untuk Bahasa Indonesia
Sigit Hanafi, Prof. Drs. Edi Winarko, M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

Pengembangan Model *Passage* Reranking Berbasis BERT Cross-Encoder untuk Bahasa Indonesia

Oleh
Sigit Hanafi
24/546999/PPA/06868

Sistem *Information Retrieval* (IR) berperan dalam menemukan dan menempatkan *passage* relevan pada peringkat atas hasil pencarian. Dalam praktiknya, BM25 masih banyak digunakan sebagai *retrieval* karena mampu melakukan pencarian secara efisien pada koleksi *passage* berskala besar. Namun, pendekatan tersebut memiliki keterbatasan dalam memahami hubungan semantik antara *query* dan *passage*. Kondisi tersebut mendorong penggunaan pendekatan *two-stage retrieval*, dengan tahap *reranking* untuk mengevaluasi kembali kandidat berdasarkan tingkat relevansinya terhadap *query* dengan mempertimbangkan hubungan kontekstual dan semantik antar teks.

Penelitian ini mengembangkan dan mengevaluasi model *reranker* berbasis BERT *cross-encoder* melalui proses *fine-tuning* dengan skema pembelajaran *pairwise* menggunakan *Margin Ranking Loss*. Dataset utama yang digunakan adalah mMARCO versi Bahasa Indonesia, sedangkan SNLI Indo digunakan sebagai data tambahan lintas domain. Eksperimen dilakukan untuk mengevaluasi pengaruh pemilihan *base model*, strategi *negative mining*, jumlah data pelatihan, serta penambahan data lintas domain terhadap performa model *reranker*. Model yang dibandingkan meliputi BERT, *multilingual* BERT, dan IndoBERT, dengan evaluasi metrik *Mean Reciprocal Rank* pada posisi 10 (MRR@10).

Hasil penelitian menunjukkan bahwa IndoBERT memperoleh performa tertinggi di antara model yang diuji pada konfigurasi penelitian dengan MRR@10 sebesar 0,2281, meningkat 48,99% dibandingkan *baseline* BM25 dengan nilai MRR@10 sebesar 0,1531. Sementara BERT dan *multilingual* BERT masing-masing memperoleh MRR@10 sebesar 0,1044 dan 0,1962. Pada eksperimen penambahan data lintas domain, seluruh variasi penambahan SNLI Indo menghasilkan performa yang lebih rendah dibandingkan model yang hanya dilatih menggunakan mMARCO. Temuan penelitian menunjukkan bahwa performa model *reranker* dipengaruhi oleh pemilihan *base model*, strategi *negative mining*, jumlah data pelatihan, serta kesesuaian karakteristik data dengan tugas utama.

Kata kunci: BERT, *cross-encoder*, *Information Retrieval*, *reranking*

UNIVERSITAS
GADJAH MADA

Pengembangan Model Passage Reranking Berbasis BERT Cross Encoder untuk Bahasa Indonesia
Sigit Hanafi, Prof. Drs. Edi Winarko, M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

Development of a BERT Cross-Encoder Model for Indonesian Passage Reranking

By
Sigit Hanafi
24/546999/PPA/06868

Information Retrieval (IR) systems play a role in retrieving and ranking relevant passages at the top of search results. In practice, BM25 is still widely used as a retrieval method due to its efficiency in searching large-scale passage collections. However, this approach has limitations in capturing semantic relationships between queries and passages. This limitation motivates the use of a two-stage retrieval approach, in which a reranking stage is employed to reassess candidate passages based on their relevance to the query by modeling contextual and semantic interactions between the query and passages.

This study develops and evaluates BERT-based cross-encoder reranker models through fine-tuning using a pairwise learning scheme with Margin Ranking Loss. The Indonesian version of the mMARCO dataset is used as the primary dataset, while SNLI Indo is incorporated as additional cross-domain data. Experiments are conducted to evaluate the effects of base model selection, negative mining strategies, training data size, and the addition of cross-domain data on reranker performance. The evaluated models include BERT, multilingual BERT, and IndoBERT, with Mean Reciprocal Rank at cutoff 10 (MRR@10) used as the evaluation metric.

The results show that IndoBERT achieves the best performance in the main training scenario, with an MRR@10 of 0.2281, representing a 48.99% improvement over the BM25 baseline of 0.1531. Meanwhile, BERT and multilingual BERT achieve MRR@10 scores of 0.1044 and 0.1962, respectively. In the cross-domain data augmentation experiments, all variations involving the addition of SNLI Indo resulted in lower performance than the model trained solely on mMARCO. These findings indicate that the performance of the reranker model is determined not only by the model architecture but also by the characteristics of the base model, the negative mining strategy, the amount of training data, and the alignment between the characteristics of the training data and the primary task.

Keywords: BERT, *reranking*, *cross-encoder*, *Information Retrieval*