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/

DAFTAR ISI

HALAMAN JUDUL ii

HALAMAN PERNYATAAN v

INTISARI xi

ABSTRACT xii

BAB I PENDAHULUAN 1

1.1 Latar Belakang Masalah 1
1.2 Rumusan Masalah 3
1.3 Batasan Masalah 3
1.4 Tujuan Penelitian 4
1.5 Manfaat Penelitian 4

BAB II TINJAUAN PUSTAKA 5

BAB III METODOLOGI PENELITIAN 10

3.1 Deskripsi Penelitian 10
3.2 Arsitektur Sistem 12
3.2.1 Arsitektur *Two-Stage Retrieval* 13
3.2.2 Model BERT *Cross-Encoder* 14
3.2.3 Mekanisme Pembelajaran Model 15
3.3 Pengambilan dan Pemuatan Dataset 17
3.3.1 Pengambilan dan Pemuatan Dataset mMARCO Bahasa Indonesia 17
3.3.2 Pengambilan dan Pemuatan Dataset SNLI Indo 19
3.4 *Preprocessing* Data 20
3.4.1 *Preprocessing* Dataset mMARCO Bahasa Indonesia 20
3.4.2 *Preprocessing* Dataset SNLI Indo 21
3.5 Pembagian Data pada mMARCO Bahasa Indonesia 22
3.6 *First-Stage Retrieval* 24
3.6.1 Metode BM25 sebagai *First-Stage Retrieval* 24
3.6.2 Proses *Indexing Passage* 25

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/

3.6.3 Perhitungan Skor BM25 26
3.6.4 Proses *Retrieval* Kandidat Top-1000 29
3.7 *Negative Mining* 30
3.7.1 Sumber Kandidat Negatif 30
3.7.2 Kategori Tingkat Kesulitan Negatif 31
3.8 Persiapan Data 34
3.8.1 Persiapan Data Pelatihan 34
3.8.2 Persiapan Data Pengujian 38
3.9 Pelatihan Model 39
3.9.1 Arsitektur *Cross-Encoder* 39
3.9.2 Pemilihan *Base Model* 42
3.9.3 Proses Pelatihan Model 43
3.9.4 Konfigurasi Pelatihan 45
3.10 Pengujian dan Evaluasi 48

BAB IV IMPLEMENTASI 51
4.1 Lingkungan Pengembangan 51
4.2 Pengambilan dan Pemuatan Dataset 52
4.3 *Preprocessing* Data 53
4.3.1 *Preprocessing* Dataset mMARCO 53
4.3.2 *Preprocessing* Dataset SNLI Indo 54
4.4 Pembentukan *Positive Pairs* 56
4.5 *First-Stage Retrieval* 57
4.5.1 Proses Indexing *Passage* 57
4.5.2 Retrieval Kandidat *Top-1000* 58
4.6 *Negative Mining* 59
4.7 Persiapan Data Pelatihan 61
4.7.1 Penyusunan *Triplet* dari Dataset mMARCO 61
4.7.2 Penyusunan *Triplet* dari Dataset SNLI Indo 62
4.7.3 Penggabungan *Triplet* mMARCO dan SNLI Indo 64
4.8 Pelatihan Model 65
4.8.1 Model Dasar dan Format Input 65
4.8.2 Konfigurasi Parameter Pelatihan 66
4.8.3 Implementasi Margin Ranking Loss 67
4.8.4 Prosedur Pelatihan Model 69

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/

4.9 Pengujian dan Evaluasi 70
4.9.1 Pemuatan Model dan Perhitungan Skor Cross-Encoder 71
4.9.2 Implementasi Perhitungan MRR 72

BAB V HASIL DAN PEMBAHASAN 74
5.1 Skenario Pengujian 74
5.2 Hasil Evaluasi *Baseline* BM25 74
5.3 Perbandingan Performa Antar Model 76
5.4 Analisis Faktor yang Mempengaruhi Performa Model 78
5.4.1 Pemilihan *Base Model* 78
5.4.2 Strategi *Negative Mining* 79
5.4.3 Jumlah Data Pelatihan 81
5.4.4 Penambahan Dataset SNLI Indo 83
5.5 Analisis Perubahan Peringkat 85
5.5.1 Analisis Kasus Peningkatan Peringkat 87
5.5.2 Analisis Kasus Penurunan Peringkat 91
5.5.3 Pola Umum Perubahan Peringkat 94
5.6 Implikasi terhadap Sistem Information Retrieval 95
5.7 Ringkasan Temuan 96

BAB VI KESIMPULAN DAN SARAN 98
6.1 Kesimpulan 98
6.2 Saran 98