

FINE-TUNING GENERAL-PURPOSE TEXT EMBEDDING (GTE)
MODEL UNTUK Pencarian Dokumen Akademik
Berbasis Semantik di Repository IPB University

HASRUL MALIK PUTRA MASHUN



PROGRAM SARJANA ILMU KOMPUTER
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025

@Hak cipta milik IPB University

IPB University



IPB University
— Bogor Indonesia —

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Perpustakaan IPB University



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “*Fine-tuning General-Purpose Text Embedding (GTE) Model* untuk Pencarian Dokumen Akademik berbasis Semantik di Repository IPB University” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apapun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Mei 2025

Hasrul Malik Putra Mashun
G6401201064

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

ABSTRAK

HASRUL MALIK PUTRA MASHUN. *Fine-Tuning General-Purpose Text Embedding (GTE) Model untuk Pencarian Dokumen Akademik berbasis Semantik di Repository IPB University*. Dibimbing oleh JULIO ADISANTOSO dan MUSHTHOFA.

Pertumbuhan jumlah dokumen di Repository IPB University menuntut sistem pencarian yang lebih relevan dan efisien, karena pencarian berbasis kata kunci sering gagal menangkap makna semantik pertanyaan pengguna. Penelitian ini bertujuan untuk meningkatkan relevansi pencarian dokumen akademik melalui *fine-tuning* embedding model GTE-Base-EN-v1.5 menggunakan pendekatan triplet dengan embedding model OpenAI. Dari total 122.977 dokumen, sebanyak 115.144 diproses menjadi 828.796 data latih dan 207.500 data validasi, dengan pelatihan menggunakan Multiple Negatives Ranking Loss. Evaluasi terhadap 740 query dari 37 departemen menunjukkan peningkatan signifikan, di mana MRR Top-5 naik dari 0.1049 menjadi 0.3938, *Precision* dari 0.0451 menjadi 0.1849, dan *Accuracy* dari 0.1554 menjadi 0.5743. Temuan ini menunjukkan bahwa pendekatan embedding berbasis NLP mampu meningkatkan kualitas sistem pencarian dalam domain akademik dan berpotensi mempercepat akses informasi yang relevan bagi civitas akademika.

Kata kunci : embedding model, fine-tuning, semantic retrieval

ABSTRACT

HASRUL MALIK PUTRA MASHUN. *Fine-Tuning a General-Purpose Text Embedding (GTE) Model for Semantic-Based Academic Document Retrieval in the IPB University Repository*. Supervised by JULIO ADISANTOSO and MUSHTHOFA.

The growing number of documents in the IPB University Repository demands a more relevant and efficient search system, as keyword-based search methods often fail to capture the semantic meaning of user queries. This study aims to improve the relevance of academic document retrieval by fine-tuning the GTE-Base-EN-v1.5 embedding model using a triplet-based approach with OpenAI embeddings model. Out of 122,977 documents, 115,144 were processed into 828,796 training data and 207,500 validation data, with training conducted using the Multiple Negatives Ranking Loss method. Evaluation on 740 queries from 37 departments showed significant improvements: Top-5 MRR increased from 0.1049 to 0.3938, Precision from 0.0451 to 0.1849, and Accuracy from 0.1554 to 0.5743. These findings demonstrate that embedding-based NLP approaches can significantly enhance the quality of academic search systems and have the potential to accelerate access to relevant information for the academic community.

Keywords: embedding model, fine-tuning, semantic retrieval



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2025
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

FINE-TUNING GENERAL-PURPOSE TEXT EMBEDDING (GTE)
MODEL UNTUK PENCARIAN DOKUMEN AKADEMIK
BERBASIS SEMANTIK DI REPOSITORY IPB UNIVERSITY

HASRUL MALIK PUTRA MASHUN

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana pada
Program Studi Ilmu Komputer

PROGRAM SARJANA ILMU KOMPUTER
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Judul

: *Fine-Tuning General-Purpose Text Embedding (GTE) Model* untuk
Pencarian Dokumen Akademik berbasis Semantik di Repository
IPB University

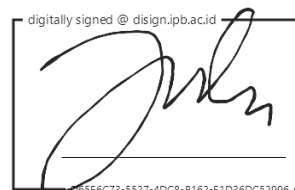
Nama
NIM

: Hasrul Malik Putra Mashun
: G6401201064

Disetujui oleh

Pembimbing 1:

Ir. Julio Adisantoso M.Kom
196207141986011002



Pembimbing 2:

Dr. Mushthofa S.Kom., M.Sc.
198203252009121003



Diketahui oleh

Ketua Program Sarjana Ilmu Komputer:

Dr. Sony Hartono Wijaya, S.Kom., M.Kom.
19810809 200812 1 002



Tanggal Ujian:

Kamis, 05 Juni 2025

Tanggal Lulus:

PRAKATA

Puji dan syukur penulis panjatkan ke hadirat Allah SWT atas segala limpahan rahmat, taufik, dan hidayah-Nya sehingga penulis dapat menyelesaikan penyusunan skripsi yang berjudul “*Fine-Tuning General-Purpose Text Embedding (GTE) Model untuk Pencarian Dokumen Akademik berbasis Semantik di Repository IPB University*” sebagai salah satu syarat untuk memperoleh gelar Sarjana di Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, IPB University.

Skripsi ini dapat terselesaikan berkat bantuan, dukungan, dan doa dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Orang tua tercinta, atas doa, kasih sayang, dan dukungan moral maupun material yang tiada henti.
2. Dosen pembimbing, Ir. Julio Adisantoso M.Kom dan Dr. Mushtofa S.Kom., M.Sc. atas bimbingan, arahan, serta masukan berharga selama proses penyusunan skripsi ini.
3. Dosen penguji dan seluruh dosen di Dean Apriana Ramadhan, S.KOM., M.Kom. atas ilmu dan wawasan yang telah diberikan selama masa perkuliahan.
4. Teman-teman seperjuangan, yang selalu memberikan semangat, diskusi, serta bantuan teknis maupun non-teknis.
5. Seluruh pihak yang tidak dapat penulis sebutkan satu per satu yang telah membantu dalam penyusunan skripsi ini.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna. Oleh karena itu, penulis sangat mengharapkan saran dan kritik yang membangun demi perbaikan di masa mendatang. Semoga skripsi ini dapat memberikan manfaat bagi pembaca dan pihak-pihak yang berkepentingan di bidang sistem pencarian dokumen akademik.

Bogor, Mei 2025

Hasrul Malik Putra Mashun

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

DAFTAR ISI

DAFTAR TABEL	iii
DAFTAR GAMBAR	iii
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Perumusan Masalah	2
1.3 Tujuan Penelitian	2
1.4 Manfaat Penelitian	2
1.5 Ruang Lingkup Penelitian	2
II TINJAUAN PUSTAKA	3
2.1 Embedding Model	3
2.2 GTE-Base-EN-v1.5	3
2.3 Semantic-Based Retrieval	4
2.4 Triplet Loss dan fine tuning embedding model	5
2.5 Evaluasi Embedding Model	6
III METODE	7
3.1 Data Penelitian	7
3.2 Tahapan Penelitian	7
3.3 Lingkungan Pengembangan Sistem	13
IV HASIL DAN PEMBAHASAN	14
4.1 Deskripsi Data Penelitian	14
4.2 Hasil Praproses Data	14
4.2 Hasil Pembentukan Embedding Vector	16
4.3 Hasil pembentukan triplet	17
4.4 Hasil Fine Tuning Model	19
4.5 Hasil Evaluasi Model	19
4.5.1 Ringkasan Hasil Evaluasi	19
4.5.2 Justifikasi Penggunaan Variasi Top-K	21
4.5.3 Analisis Detail per Metrik Evaluasi	22
4.5.4 Implikasi dan Signifikansi Hasil	27
4.5.5 Limitasi dan Pertimbangan	28
4.6 Analisis Studi Kasus Query Spesifik	28
4.7 Pengembangan Sistem di Masa Depan	31
V KESIMPULAN DAN SARAN	32
5.1 Kesimpulan	32
5.2 Saran	32
DAFTAR PUSTAKA	33
RIWAYAT HIDUP	35

DAFTAR TABEL

1	Contoh data setelah dilakukan <i>generate embedding vector</i>	9
2	Contoh data triplet	10
3	Pola regex untuk penyederhanaan kolom di dataset	15
4	Data bersih hasil akhir pra-proses data	16
5	Data hasil pembentukan <i>embedding vector</i>	17
6	Data triplet	18
7	Perbandingan performa model sebelum dan setelah <i>fine-tuning</i>	20
8	Perbandingan MRR sebelum dan sesudah <i>fine-tuning</i>	22
9	Perbandingan <i>precision</i> sebelum dan sesudah <i>fine-tuning</i>	23
10	Perbandingan <i>recall</i> sebelum dan sesudah <i>fine-tuning</i>	24
11	Perbandingan <i>accuracy</i> sebelum dan sesudah <i>fine-tuning</i>	25
12	Perbandingan NDCG sebelum dan sesudah <i>fine-tuning</i>	26

DAFTAR GAMBAR

1	Diagram alur penelitian	7
2	Grafik tren <i>training loss fine-tuning</i>	19
3	Plot perbandingan MRR	22
4	Plot perbandingan <i>precision</i>	23
5	Plot perbandingan <i>recall</i>	24
6	Plot perbandingan <i>accuracy</i>	25
7	Plot perbandingan NDCG	26