



PENERAPAN *RETRIEVE-AND-RERANK ENTITY RESOLUTION* UNTUK DETEKSI DUPLIKAT NONIDENTIK DATA AKTIVITAS MAHASISWA IPB

MUHAMMAD DZAKWAN ALIFI



**PROGRAM STUDI SARJANA STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA MATEMATIKA DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**



@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.



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Penerapan *Retrieve-and-Rerank Entity Resolution* untuk Deteksi Duplikat Nonidentik Data Aktivitas Mahasiswa IPB” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun 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.

Dalam penyusunan karya ini, penulis menggunakan bantuan kecerdasan buatan Gemini 3.1 Flash-Lite untuk verifikasi pasangan data dan Nano Banana 2 untuk pembuatan ilustrasi/visualisasi gambar. Setelah menggunakan alat/layanan tersebut, penulis meninjau dan menyunting konten sesuai kebutuhan serta bertanggung jawab penuh atas isi karya tugas akhir ini.

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

Bogor, Juni 2026

Muhammad Dzakwan Alifi
G1401221049



ABSTRAK

MUHAMMAD DZAKWAN ALIFI. Penerapan *Retrieve-and-Rerank Entity Resolution* untuk Deteksi Duplikat Nonidentik Data Aktivitas Mahasiswa IPB. Dibimbing oleh RAHMA ANISA dan GERRY ALFA DITO.

Data aktivitas mahasiswa IPB University periode 2019/2020–2025/2026 memiliki variasi penulisan yang tinggi, seperti penggunaan akronim yang tidak seragam, kesalahan pengetikan, serta penggunaan istilah campuran bahasa. Hal tersebut menyebabkan duplikat nonidentik sulit dideteksi oleh metode berbasis pencocokan *string literal* seperti *Exact Match* dan *Fuzzy Match*. Penelitian ini bertujuan mengadaptasi dan mengevaluasi metode *entity resolution* EnsembleLink berbasis *retrieve-and-rerank* untuk deteksi dan pengelompokan duplikat nonidentik *intradataset* pada data aktivitas mahasiswa IPB secara *zero-shot*. Metode yang diadaptasi memadukan *sparse retrieval* TF-IDF, *dense retrieval* IndoE5, penyaringan *cross-encoder*, dan pengelompokan *connected components*. Evaluasi kinerja dilakukan terhadap data acuan yang disusun dengan bantuan LLM dan divalidasi secara sampling untuk memastikan kesesuaian pasangan serta kluster duplikat nonidentik. Hasil evaluasi menunjukkan bahwa EnsembleLink mencapai *F1-score* 0,7892, *recall* 0,8447, dan *precision* 0,7405 pada ambang batas 0,7, yang mengungguli metode pencocokan konvensional. Modul *reranker* terbukti efektif mencegah kesalahan penggabungan kluster. Penerapan metode ini dapat mendukung pembersihan data aktivitas kemahasiswaan secara semi-otomatis pada tingkat institusi.

Kata kunci: aktivitas mahasiswa, duplikat nonidentik, *entity resolution*, *retrieve-and-rerank*, *zero-shot learning*

@Hak Cipta: <https://doi.org/10.24127/ijer.v1i1.12345>

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 mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

ABSTRACT

MUHAMMAD DZAKWAN ALIFI. Application of Retrieve-and-Rerank Entity Resolution for Non-Identical Duplicate Detection in IPB Student Activity Data. Supervised by RAHMA ANISA and GERRY ALFA DITO.

IPB University student activity records from 2019/2020–2025/2026 exhibited high spelling variations, such as non-uniform acronym usage, typographical errors, and mixed-language terms. These variations made non-identical duplicates difficult to detect using string-literal matching methods such as Exact Match and Fuzzy Match. This study aimed to adapt and evaluate the EnsembleLink retrieve-and-rerank entity resolution framework for zero-shot detection and clustering of non-identical duplicates within the IPB student activity dataset. The adapted method integrated TF-IDF sparse retrieval, IndoE5 dense retrieval, cross-encoder filtering, and connected components clustering. Performance evaluation was conducted against a ground truth dataset constructed with LLM assistance and validated via sampling to ensure the correctness of pairs and non-identical duplicate clusters. Evaluation results showed that EnsembleLink achieved an F1-score of 0.7892, recall of 0.8447, and precision of 0.7405 at a 0.7 threshold, which outperformed conventional matching methods. The reranker module was proven to effectively prevent erroneous cluster merges. The implementation of this method supported semi-automatic student activity data cleaning at the institutional level.

Keywords: entity resolution, non-identical duplicates, retrieve-and-rerank, student activities, zero-shot learning



@Hak cipta milik IPB University

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.

© Hak Cipta milik IPB, tahun 2026
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 apa pun tanpa izin IPB.



**PENERAPAN *RETRIEVE-AND-RERANK ENTITY RESOLUTION*
UNTUK DETEKSI DUPLIKAT NONIDENTIK
DATA AKTIVITAS MAHASISWA IPB**

MUHAMMAD DZAKWAN ALIFI

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana pada
Program Studi Statistika dan Sains Data

**PROGRAM STUDI SARJANA STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA MATEMATIKA DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**



@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.

Tim Penguji pada Ujian Skripsi:
Pika Silvianti S.Si., M.Si.



@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 Skripsi : Penerapan *Retrieve-and-Rerank Entity Resolution* untuk
Deteksi Duplikat Nonidentik Data Aktivitas Mahasiswa IPB

Nama : Muhammad Dzakwan Alifi
NIM : G1401221049

@Hak cipta milik IPB University

Disetujui oleh

Pembimbing 1:
Rahma Anisa S.Stat., M.Si., M.Act.Sc.

Pembimbing 2:
Gerry Alfa Dito S.Si., M.Si.

Diketahui oleh

Ketua Program Studi:
Dr. Bagus Sartono, S.Si, M.Si
NIP 19780411 200501 1002

Tanggal Ujian:
2 Juni 2026

Tanggal Lulus:

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Penelitian yang dilaksanakan sejak bulan Oktober 2025 sampai bulan April 2026 ini berfokus pada topik resolusi entitas (*entity resolution*), dengan judul “Penerapan *Retrieve-and-Rerank Entity Resolution* untuk Deteksi Duplikat Nonidentik Data Aktivitas Mahasiswa IPB”. Ucapan terima kasih penulis sampaikan kepada:

1. Ibu Rahma Anisa, S.Stat., M.Si., M.Act.Sc. dan Bapak Gerry Alfa Dito, S.Si., M.Si. selaku komisi pembimbing yang telah membimbing dan banyak memberi saran;
2. pembimbing akademik, moderator seminar proposal Ibu Eka Putri Nur Utami, S.Stat., M.Si., moderator seminar hasil Bapak Dr. Farit Mochamad Afendi, S.Si., M.Si., serta penguji luar komisi pembimbing Ibu Pika Silvianti, S.Si., M.Si. atas masukan dan bimbingan yang diberikan;
3. Bapak M. Faturrokman, S.Pt., M.Si. selaku Asisten Direktur Pembinaan Karakter, Ormawa, dan Orsen, Direktorat Kemahasiswaan IPB University, serta Saudara Ghiffari selaku *project lead* pengembangan sistem yang telah mengizinkan penggunaan data aktivitas mahasiswa untuk penelitian ini;
4. Bapak Almaskur dan Ibu Indra Purnamasari, adik-adik penulis yaitu Nafis dan Alya, serta seluruh keluarga yang telah memberikan dukungan, doa, dan kasih sayangnya;
5. Saudara Rakesha selaku Ketua Umum Himpunan Profesi Statistika Gamma Sigma Beta (GSB) IPB University periode 2024/2025, BPH, seluruh pengurus, serta Departemen *Data Analytics* periode 2022/2023 dan 2023/2024 selaku staf magang dan staf departemen atas kerja samanya;
6. teman-teman angkatan 59 Statistika (Marinestic) yang telah menemani perjuangan selama masa perkuliahan;
7. teman-teman Mahasiswa Berprestasi IPB University tahun 2025 dan Kaylah atas lingkungan yang suportif, kompetitif, dan prestatif yang mendorong penulis untuk terus berkembang; serta
8. teman-teman Rumah Oren (Alfan, Fatih, Firhan, dan Auzan) atas keceriaan dan kebersamaan hangat yang menemani perjalanan perkuliahan selama tiga tahun terakhir.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Juni 2026

Muhammad Dzakwan Alifi



DAFTAR ISI

DAFTAR TABEL	xi
DAFTAR GAMBAR	xii
DAFTAR LAMPIRAN	xii
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	3
TINJAUAN PUSTAKA	4
2.1 Sistem Informasi Kemahasiswaan	4
2.2 <i>Entity Resolution</i> dan <i>Record Linkage</i>	5
2.3 <i>Dense Retrieval</i> dengan IndoE5	6
2.4 <i>Sparse Retrieval</i> dengan TF-IDF <i>Character N-gram</i>	7
2.5 <i>Hybrid Retrieval</i> Gabungan <i>Dense</i> dan <i>Sparse</i>	9
2.6 <i>Cross-Encoder Reranking</i>	10
2.7 Anotasi Data dengan <i>Large Language Model (LLM)</i>	11
2.8 Pengelompokan Berbasis Graf dengan <i>Connected Components</i>	12
2.9 Metrik Evaluasi <i>Entity Resolution</i>	14
III METODE	16
3.1 Sumber dan Karakteristik Data	16
3.2 Prosedur Penelitian	18
3.3 Analisis Data	27
3.4 Spesifikasi Lingkungan Komputasi	28
IV HASIL DAN PEMBAHASAN	29
4.1 Hasil Pembersihan Data dan Karakteristik Data Acuan	29
4.2 Evaluasi Perbandingan Metode	32
4.3 Studi Ablasi Komponen <i>Pipeline</i>	37
4.4 Analisis Sensitivitas Ambang Batas	41
4.5 Analisis Kesalahan	42
4.6 Perbandingan Waktu Komputasi	46
V SIMPULAN DAN SARAN	48
5.1 Simpulan	48
5.2 Saran	48
DAFTAR PUSTAKA	49
LAMPIRAN	52
RIWAYAT HIDUP	58

DAFTAR TABEL

1	Sampel variasi penulisan aktivitas mahasiswa beserta kategori variasinya	16
2	Contoh nilai <i>cosine similarity dense</i> antarpasangan teks	22
3	Contoh nilai <i>cosine similarity sparse</i> antarpasangan teks	23
4	Contoh skor <i>cross-encoder</i> pada pasangan data aktivitas mahasiswa	24
5	Ringkasan metode <i>baseline</i> yang digunakan sebagai pembandingan	26
6	Konfigurasi studi ablasi	27
7	Pustaka utama yang digunakan dalam penelitian beserta fungsinya	28
8	Tahapan pembersihan data dan jumlah baris yang dihapus per prosedur	29
9	Sampel data aktivitas yang dieliminasi berdasarkan kriteria penyaringan <i>noise</i> dan anomali karakter	29
10	Distribusi ukuran keanggotaan klaster pada data acuan	30
11	Contoh hasil verifikasi kesamaan antaraktivitas oleh LLM	31
12	Contoh variasi penulisan aktivitas mahasiswa dalam klaster data acuan	31
13	Perbandingan kinerja seluruh metode deteksi duplikat nonidentik	32
14	Perbandingan cakupan <i>retrieval</i> antarmetode pada contoh konkret dari data penelitian	34
15	Dekomposisi sumber penemuan kandidat data duplikat nonidentik pada <i>hybrid retrieval</i>	35
16	Hasil studi ablasi komponen EnsembleLink	38
17	Contoh rantai <i>chaining effect</i> pada konfigurasi tanpa <i>reranker</i>	38
18	Contoh skor <i>cross-encoder reranker</i> pada dua kueri representatif	39
19	Sensitivitas EnsembleLink terhadap ambang batas <i>cross-encoder</i> (τ)	41
20	Contoh pasangan yang keliru digabungkan (<i>over-merged</i>) pada $\tau = 0,5$	43
21	Contoh pasangan duplikat nonidentik yang gagal disatukan oleh EnsembleLink (<i>under-merged</i>)	44
22	Perbandingan perkiraan waktu komputasi dan <i>F1-score</i>	46

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 mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



DAFTAR GAMBAR

1	Arsitektur SBERT saat pelatihan (a) dan saat inferensi (b) (Reimers dan Gurevych 2019).	7
2	Empat arsitektur <i>retrieval</i> : (a) <i>cross-encoder</i> murni, (b) <i>bi-encoder</i> , (c) <i>cooperative retrieve-and-rerank</i> (SEP+COOP), (d) <i>joint retrieve-and-rerank</i> (JOINT+COOP) (Geigle et al. 2022).	10
3	Ilustrasi tiga algoritme <i>single-pass clustering</i> pada graf kemiripan yang sama: (a) <i>Partitioning/connected components</i> , (b) <i>CENTER</i> , (c) <i>MERGE-CENTER</i> (Hassanzadeh et al. 2009).	13
4	Diagram alur arsitektur <i>pipeline</i> EnsembleLink yang diadaptasi	18
5	Ilustrasi proses <i>dense retrieval</i> IndoE5, konversi teks aktivitas menjadi vektor padat dan representasinya dalam ruang vektor (dihasilkan oleh Nano Banana 2 [Google 2026])	21
6	Ilustrasi <i>sparse retrieval</i> TF-IDF <i>character n-gram</i> pada teks (dihasilkan oleh Nano Banana 2 [Google 2026])	23
7	Pembentukan kluster pasangan duplikat nonidentik menggunakan sifat transitif (dihasilkan oleh Nano Banana 2 [Google 2026])	26
8	Dekomposisi sumber penemuan kandidat duplikat nonidentik pada 59.916 pasangan data acuan	37
9	Ilustrasi <i>chaining effect</i> yang mengaitkan dua aktivitas tidak berhubungan melalui satu simpul mediasi	38
10	Kurva <i>trade-off precision</i> dan <i>recall</i> pada variasi ambang batas τ	42

DAFTAR LAMPIRAN

1	Lampiran 1 Sistem Prompt LLM Anotator <i>Entity Matching</i>	53
2	Lampiran 2 Kode Program Utama <i>Pipeline</i> EnsembleLink	54
3	Lampiran 3 Representasi Matriks <i>Dense Embedding</i> IndoE5	55
4	Lampiran 4 Representasi Matriks <i>Sparse TF-IDF Character N-gram</i>	56
5	Lampiran 5 Representasi Matriks Skor <i>Cross-Encoder Reranker</i>	57