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

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 klaster 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 klaster. 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*

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

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

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

Judul Skripsi : Penerapan *Retrieve-and-Rerank Entity Resolution* untuk Deteksi Duplikat Nonidentik Data Aktivitas Mahasiswa IPB
Nama : Muhammad Dzakwan Alifi
NIM : G1401221049

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-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

DAFTAR GAMBAR

DAFTAR LAMPIRAN