# ABSTRAK

## PENGEMBANGAN SISTEM PENDETEKSI KESAMAAN KONTEKS PADA DOKUMEN AKADEMIK MAHASISWA JURUSAN ILMU KOMPUTER UNIVERSITAS LAMPUNG

**Oleh**

**Fitria Az Zahra**

Kemiripan dokumen dan plagiarisme masih menjadi permasalahan yang sering dijumpai dalam penyusunan tugas akhir mahasiswa, termasuk di Jurusan Ilmu Komputer Universitas Lampung. Sistem perpustakaan digital yang tersedia umumnya masih mengandalkan pencarian berbasis kata kunci sehingga kurang mampu mengidentifikasi kesamaan makna antardokumen. Untuk mengatasi permasalahan tersebut, penelitian ini mengembangkan SimTA, yaitu sistem deteksi kemiripan dokumen berbasis web yang dibangun menggunakan *framework* Flask. Sistem ini menerapkan pendekatan *hybrid scoring* yang menggabungkan analisis leksikal menggunakan TF-IDF dan analisis semantik menggunakan *Sentence*-BERT (SBERT) dengan model pra-latih paraphrase-multilingual-MiniLM-L12-v2. SimTA mendukung unggah dokumen dalam format PDF, DOCX, dan TXT serta menghitung tingkat kemiripan menggunakan *Cosine Similarity*. Evaluasi dilakukan terhadap 140 dokumen tugas akhir mahasiswa yang terdiri dari 30 dokumen D3 Manajemen Informatika dan 110 dokumen skripsi S1 Ilmu Komputer. Kinerja sistem dievaluasi menggunakan metrik *Precision, Recall,* dan *Mean Average Precision* (MAP). Hasil pengujian menunjukkan bahwa model *hybrid* memberikan performa terbaik dengan nilai *Precision* sebesar 90%, *Recall* sebesar 93,8%, dan MAP sebesar 91,5%, lebih tinggi dibandingkan model TF-IDF yang memperoleh MAP 85,5% dan model SBERT dengan MAP 80,9%. Penemuan ini menunjukkan bahwa kombinasi pendekatan berbasis kata dan makna mampu meningkatkan akurasi dalam mendeteksi kemiripan dokumen tugas akhir mahasiswa.

**Kata Kunci:** *Sentence*-BERT, TF-IDF, *Cosine Similarity*, Kemiripan Dokumen Pemrosesan Bahasa Alami.

## *ABSTRACT*

## *DEVELOPMENT OF A CONTEXTUAL SIMILARITY DETECTION SYSTEM FOR STUDENT ACADEMIC DOCUMENTS IN THE DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITAS LAMPUNG*

***By***

**Fitria Az Zahra**

*Document similarity and plagiarism remain common challenges in the preparation of student final project documents, including in the Department of Computer Science at Universitas Lampung. Existing digital library systems generally rely on keyword-based retrieval, which limits their ability to effectively identify semantic similarities between documents. To address this issue, this study developed SimTA, a web-based document similarity detection system built using the Flask framework. The system employs a hybrid scoring approach that combines lexical analysis using TF-IDF and semantic analysis using Sentence-BERT (SBERT) with the pre-trained paraphrase-multilingual-MiniLM-L12-v2 model. SimTA supports document uploads in PDF, DOCX, and TXT formats and calculates similarity scores using Cosine Similarity. The evaluation was conducted on 140 student final project documents consisting of 30 final project reports from the D3 Informatics Management program and 110 undergraduate theses from the Computer Science program. System performance was evaluated using Precision, Recall, and Mean Average Precision (MAP). The experimental results show that the hybrid model achieved the best performance, with a Precision of 90%, Recall of 93.8%, and MAP of 91.5%, outperforming the TF-IDF model, which achieved a MAP of 85.5%, and the SBERT model, which obtained a MAP of 80.9%. These findings indicate that combining keyword-based and semantic-based approaches can improve the accuracy of detecting similarities in student final project documents.*



***Keywords:*** *Sentence-BERT;* TF-IDF*; Cosine Similarity; Document Similarity; Natural Language Processing*