

ABSTRAK

TAWAKKAL MA'MUN, Integrasi *Google CSE* dan *Sentence-BERT* untuk deteksi Plagiarisme Semantik Pada Skripsi Mahasiswa Teknik UNISMUH Makassar (dibimbing oleh Ir. Ida, S.Kom.,M.T, dan Chyquitha Danuputri, S.Kom.,M.Kom).

Plagiarisme semantik pada skripsi, terutama dalam bentuk parafrase yang mempertahankan makna tetapi mengubah susunan kalimat, masih sulit dideteksi oleh metode *konvensional* berbasis *string matching* seperti yang umum digunakan di perguruan tinggi. Penelitian ini bertujuan mengembangkan sistem deteksi plagiarisme semantik pada skripsi mahasiswa Teknik Informatika UNISMUH Makassar melalui integrasi *Google Custom Search Engine (CSE)*, korpus lokal skripsi, dan model *Sentence-BERT (SBERT)* dalam satu alur kerja *hybrid* per segmen. Dokumen skripsi diunggah melalui antarmuka web, diekstraksi teksnya, lalu dipotong menjadi segmen berukuran 25 kata dengan overlap 5 kata sebelum diubah menjadi *vektor embedding* berdimensi 768 menggunakan model *paraphrase-multilingual-mpnet-base-v2*. Setiap segmen kemudian dibandingkan secara semantik dengan kandidat referensi dari dua sumber (*Google CSE* dan korpus lokal) menggunakan *cosine similarity*, dan diberi label Plagiat atau Original berdasarkan ambang batas 0,75, sementara pengujian fungsional dilakukan dengan metode *black box testing*. Hasil pengujian menunjukkan bahwa sistem dapat mendeteksi indikasi plagiarisme semantik pada 38 dari 172 segmen (22,09%) dengan rata-rata *cosine similarity* 63,22 dan waktu proses sekitar 76,2 detik, sedangkan pemeriksaan dokumen yang sama menggunakan Turnitin hanya menghasilkan *similarity index* global sebesar 6% dan dinyatakan aman. Temuan ini menunjukkan bahwa pendekatan deteksi semantik per segmen berbasis *SBERT* lebih sensitif dalam mengidentifikasi parafrase dan kesamaan ide dibanding metode *konvensional* yang hanya memberikan satu skor persentase pada tingkat dokumen, sehingga berpotensi menjadi alat bantu yang lebih efektif dalam menjaga integritas akademik di lingkungan UNISMUH Makassar.

Kata kunci: plagiarisme semantik, *Sentence-BERT*, *Google Custom Search Engine*, *cosine similarity*, skripsi.

ABSTRACT

TAWAKKAL MA'MUN, Integration of *Google CSE* and *Sentence-BERT* for Semantic Plagiarism Detection in Theses of Engineering Students at UNISMUH Makassar

(Supervised by Ir. Ida, S.Kom., M.T. and Chyquitha Danuputri, S.Kom., M.Kom.).

Semantic plagiarism in theses, particularly in the form of paraphrasing that retains meaning while altering Sentence structure, remains difficult to detect using conventional string-matching methods commonly employed at universities. This research aims to develop a semantic plagiarism detection system for theses of Informatics Engineering students at UNISMUH Makassar through the integration of Google Custom Search Engine (CSE), local thesis corpus, and Sentence-BERT (SBERT) model within a hybrid per-segment workflow. The thesis document is uploaded via web interface, with text extracted and segmented into 25-word chunks with 5-word overlap before being converted into 768-dimensional embedding vectors using the paraphrase-multilingual-mpnet-base-v2 model. Each segment is then semantically compared against reference candidates from dual sources (Google CSE and local corpus) using cosine similarity, and labeled as "Plagiat" or "Original" based on a 0.75 threshold, while functional testing is conducted through black box methodology. Testing results show the system detecting semantic plagiarism indications in 38 out of 172 segments (22.09%) with an average cosine similarity of 63.22 and processing time of approximately 76.2 seconds, whereas examination of the same document using Turnitin yielded only a 6% global similarity index deemed safe. These findings demonstrate that the SBERT-based per-segment semantic detection approach is more sensitive in identifying paraphrases and conceptual similarities compared to conventional methods that provide only a single document-level percentage score, positioning it as a potentially more effective tool for maintaining academic integrity within the UNISMUH Makassar environment

Keywords: *semantic plagiarism, Sentence-BERT, Google Custom Search Engine, cosine similarity, thesis.*