BAB I
PENDAHULUAN

1.1 Latar Belakang

Integritas akademik di era digital kini menghadapi tantangan serius seiring semakin mudahnya akses informasi dan pertukaran konten. Dalam konteks ini, praktik plagiarisme telah berevolusi dari penyalinan verbatim menjadi modus yang lebih halus seperti parafrase, yaitu mengganti susunan kalimat atau sinonim tanpa mengubah substansi makna [18]. Fenomena ini bukan sekadar masalah administrasi, melainkan krisis nilai fundamental di lingkungan pendidikan tinggi yang semakin mendesak di era generatif AI [14]. Kesulitan utama pada modus parafrase adalah kemiripan semantik dapat tetap tinggi meskipun kemiripan leksikal (kata per kata) rendah, sehingga pemeriksaan yang hanya bertumpu pada teks permukaan sering luput mendeteksi [17].

Pada praktiknya, verifikasi manual membutuhkan waktu yang lama dan tidak scalable. Di sisi lain, pendekatan komputasi yang hanya mengandalkan text preprocessing standar tidak selalu efektif, karena keputusan dalam tahap pembersihan teks dapat memengaruhi representasi makna secara drastis [1]. Oleh sebab itu, strategi deteksi plagiarisme modern menuntut alur kerja yang lebih komprehensif [8]. Keterbatasan instrumen deteksi konvensional semakin nyata karena mayoritas masih bersifat leksikal. Metode seperti Regular Expression memang efektif untuk normalisasi pola, namun tetap bertumpu pada bentuk fisik teks [4]. Pada teks akademik pendek, algoritma String Matching klasik sensitif terhadap perubahan kecil dan kurang representatif untuk parafrase [13]. Sementara itu, metode TF-IDF yang dipadukan dengan Cosine Similarity masih menjadi standar baseline pada dokumen berbahasa Indonesia, termasuk tesis [2], sistem temu-kembali informasi di repositori kampus [5], hingga deteksi kemiripan source code [10]. Secara umum, pendekatan leksikal ini tetap populer untuk dokumen elektronik karena implementasinya yang cepat [19]. Namun, performa pendekatan leksikal terbukti mengalami degradasi signifikan ketika berhadapan dengan manipulasi sinonim, karena model ini tidak memahami konteks makna dan relasi semantik antar kata [12].

Untuk mengatasi stagnasi akurasi tersebut, riset terkini mengarah pada penggabungan sinyal leksikal dan semantik, misalnya melalui ensemble TF-IDF dengan embedding berbasis BERT [3]. Studi komparasi menunjukkan bahwa representasi vektor dokumen (Doc2Vec) lebih representatif dibanding vektor kata (Word2Vec) [6]. Secara spesifik, pendekatan Sentence-BERT telah terbukti efektif untuk pencarian semantik pada skripsi berformat PDF [9]. Namun, mayoritas penelitian tersebut masih berfokus pada Bahasa Inggris atau menggunakan model Transformer berukuran besar yang berat secara komputasi. Belum banyak penelitian yang menguji efektivitas model Transformer ringan (MiniLM) secara spesifik pada karakteristik tugas mahasiswa Teknik Informatika di Indonesia yang kaya akan istilah teknis.

Berdasarkan kesenjangan tersebut, penelitian ini berfokus pada pembangunan sistem deteksi plagiarisme yang efisien dan akurat. Sistem dibangun menggunakan metode Sentence

Embedding (all-MiniLM-L6-v2) untuk merepresentasikan makna kalimat, kemudian diukur menggunakan algoritma Cosine Similarity. Pendekatan ini dievaluasi secara komparatif terhadap baseline TF-IDF untuk membuktikan efektivitasnya dalam mendeteksi plagiarisme tipe parafrase (plagiarisme ide) pada dokumen akademik mahasiswa.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, inti permasalahan yang akan dikaji dalam penelitian ini dirumuskan sebagai berikut:

1. Bagaimana kinerja metode *Sentence Embedding* berbasis *Transformer* dalam merepresentasikan semantik kalimat pada dokumen tugas mahasiswa yang memiliki variasi struktur bahasa (parafrase)?
2. Seberapa besar tingkat akurasi deteksi plagiarisme yang dihasilkan oleh algoritma *Cosine Similarity* ketika diterapkan pada vektor semantik dibandingkan dengan metode konvensional?
3. Bagaimana hasil analisis komparatif antara pendekatan *Sentence Embedding* dan metode konvensional dalam mengidentifikasi tingkat kemiripan dokumen, serta metode manakah yang lebih efektif untuk kasus plagiarisme ide?

1.3 Batasan Masalah

Agar penelitian ini tetap terarah dan tidak meluas dari fokus utama, penulis menetapkan batasan-batasan masalah sebagai berikut:

1. Objek Penelitian: Data yang digunakan ditingkatkan menjadi 600 baris data yang disusun dalam bentuk pasangan teks (text pairs). Sumber data berasal dari repositori dokumen tugas mahasiswa Program Studi Teknik Informatika Universitas Prima Indonesia dalam format teks digital (.docx, .pdf, atau .txt). Data ini telah dianotasi secara manual dengan label Ground Truth (0 atau 1) sebagai acuan kebenaran.
2. Metode Utama (Model Bahasa): Sistem mengimplementasikan metode Sentence Embedding dengan membandingkan dua model pre-trained Transformer, yaitu:
    - all-MiniLM-L6-v2 (Model ringan dan cepat).
    - paraphrase-multilingual-MiniLM-L12-v2 (Model multilingual yang direkomendasikan untuk performa analitis teks yang lebih baik)
3. Metode Pembanding: Metode konvensional yang digunakan sebagai tolak ukur (*baseline*) adalah pembobotan TF-IDF (*Term Frequency-Inverse Document Frequency*).
4. Algoritma Pengukuran: Perhitungan tingkat kemiripan antar dokumen menggunakan algoritma *Cosine Similarity*.
5. Lingkup Deteksi: Penelitian ini berfokus pada deteksi plagiarisme teks dan parafrase, tidak mencakup deteksi pada elemen gambar, rumus matematika kompleks, atau kode program.

1.4 Tujuan Penelitian

Mengacu pada rumusan masalah di atas, tujuan spesifik yang ingin dicapai dalam penelitian ini adalah:

1. Mengimplementasikan dan menguji kinerja model *Sentence Embedding* berbasis *Transformer* dalam menangkap makna semantik pada dokumen tugas mahasiswa.
2. Membandingkan performa antara model all-MiniLM-L6-v2 dan paraphrase-multilingual-MiniLM-L12-v2 untuk menemukan model paling optimal dalam menangani teks Bahasa Indonesia.
3. Mengukur tingkat akurasi dan menentukan nilai ambang batas (*threshold*) yang optimal untuk deteksi plagiarisme menggunakan metode *Cosine Similarity* berbasis vektor semantik.
4. Menganalisis perbandingan performa antara metode yang diusulkan (Sentence Embedding) dengan metode konvensional (TF-IDF) untuk membuktikan efektivitas sistem dalam menangani kasus plagiarisme parafrase.

1.5 Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat sebagai berikut:

1. Manfaat Teoretis:
    - Memberikan wawasan akademis mengenai penerapan teknologi *Natural Language Processing* (NLP) modern, khususnya efektivitas model *Transformer* dalam memproses teks Bahasa Indonesia.
    - Menjadi referensi bagi penelitian selanjutnya yang berfokus pada pengembangan sistem deteksi kesamaan semantik (*Semantic Textual Similarity*).
2. Manfaat Praktis:
    - Bagi Dosen dan Institusi: Menyediakan alat bantu verifikasi orisinalitas tugas yang lebih cerdas dan efisien, sehingga dapat meningkatkan integritas akademik dan meminimalisir kecurangan yang sulit dideteksi mata telanjang.
    - Bagi Mahasiswa: Mendorong budaya kejujuran akademik dan orisinalitas dalam berkarya, karena sistem mampu mendeteksi praktik penyalinan ide meskipun struktur kalimat telah diubah.

1.6 Keterbaruan Penelitian

Penelitian mengenai deteksi plagiarisme telah banyak dilakukan sebelumnya. Namun, penelitian ini memiliki aspek keterbaruan (novelty) yang membedakannya dari studi terdahulu, yang dirangkum sebagai berikut:

1. Analisis Komparatif Multi-Model (Transformer vs Baseline):

Penelitian ini tidak hanya membandingkan pendekatan semantik dengan leksikal (TF-IDF), tetapi juga melakukan komparasi internal antara dua varian model Transformer, yaitu model ringan (all-MiniLM-L6-v2) dengan model multilingual kompleks (paraphrase-multilingual-MiniLM-L12-v2). Hal ini memberikan wawasan baru mengenai trade-off antara efisiensi komputasi dan akurasi semantik pada bahasa lokal.

2. Keterbaruan Objek dan Konteks (Domain-Specific):

   Penelitian ini menyajikan evaluasi spesifik pada 600 pasangan dokumen tugas mahasiswa Teknik Informatika Universitas Prima Indonesia. Hal ini melengkapi penelitian terdahulu yang mayoritas berfokus pada kemiripan judul skripsi [15] atau ekstraksi kata kunci [7], dengan memperluas cakupan analisis ke level kalimat penuh yang memiliki terminologi teknis spesifik.

3. Optimalisasi Threshold Berbasis Ground Truth:

   Penelitian ini memberikan kontribusi berupa penentuan nilai ambang batas (threshold) kemiripan yang divalidasi secara langsung menggunakan kunci jawaban manual (Ground Truth) berskala 0-1. Karakteristik dokumen akademik membutuhkan kebijakan ambang yang presisi agar interpretasi sistem (Plagiat/Tidak) menjadi valid dan terukur secara ilmiah [20], [11].