Deteksi Plagiarisme Jawaban Antar Siswa Berbasis Google Form Menggunakan *BERT* dan *Cosine Similarity*

Talitha Nadia Zulfa Irwan[1], Anita Qoiriah[2]

[1,2] Program Studi S1 Teknik Informatika, Universitas Negeri Surabaya
[1]talithanadia.22011@mhs.unesa.ac.id
[2]anitaqoiriah@unesa.ac.id

***Abstrak*— Plagiarisme antar siswa pada jawaban esai masih menjadi tantangan dalam pembelajaran daring, terutama pada pengumpulan jawaban melalui Google Form. Proses pemeriksaan kemiripan jawaban secara manual memerlukan waktu yang lama dan sulit untuk mengidentifikasi kemiripan semantik antarjawaban siswa. Penelitian ini bertujuan untuk mengembangkan sistem deteksi plagiarisme jawaban antar siswa menggunakan Sentence-BERT (*all-MiniLM-L12-v2*) dan *cosine similarity*. Tahapan penelitian meliputi *text preprocessing*, pembentukan *sentence embedding*, perhitungan skor kemiripan, penentuan batas ambang (*threshold*), serta klasifikasi jawaban ke dalam kategori plagiarisme, parafrase, dan tidak plagiarisme. Pelabelan manual dilakukan oleh dua penilai, yaitu peneliti dan guru mata pelajaran, untuk menetapkan *ground truth*. Tingkat kesepakatan antara kedua penilai kemudian dievaluasi menggunakan Cohen's Kappa yang menghasilkan nilai 0,8954 menunjukkan tingkat kesepakatan hampir sempurna. Berdasarkan evaluasi menggunakan confusion matrix terhadap empat skenario threshold, skenario 2 memberikan performa terbaik dengan akurasi sebesar 94,42%. Skenario tersebut menggunakan batas ambang similarity ≥80% untuk kategori plagiarisme, 60%–79% untuk kategori parafrase, dan <60% untuk kategori tidak plagiarisme. Nilai F1-score yang diperoleh masing-masing sebesar 0,99 pada kategori plagiarisme, 0,93 pada kategori parafrase, dan 0,88 pada kategori tidak plagiarisme.**

Kata Kunci— Plagiarisme, Sentence-BERT, *cosine similarity*, Google Form, NLP.

I. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi telah memberikan perubahan besar dalam berbagai aspek kehidupan manusia, termasuk dalam bidang pendidikan [1]. Pemanfaatan teknologi dalam proses pembelajaran saat ini tidak hanya digunakan sebagai media penyampaian materi, tetapi juga mendukung proses pemberian tugas dan pengumpulan tugas secara daring. Pada jenjang Sekolah Menengah Atas (SMA), Google Form masih menjadi platform yang banyak dimanfaatkan oleh guru karena memungkinkan siswa mengerjakan tugas secara *daring* melalui tautan yang dapat diakses menggunakan laptop atau smartphone. Selain itu, jawaban siswa akan tersimpan secara *real-time* dalam *spreadsheet* sehingga memudahkan guru dalam mengelola hasil pengumpulan dari jawaban siswa[2].

Selain memberikan kemudahan, penggunaan Google Form juga memiliki tantangan yaitu meningkatkan potensi terjadinya plagiarisme antar siswa yang dilakukan secara sengaja maupun tidak sengaja [3]. Hal ini dipengaruhi oleh kemudahan akses informasi melalui internet serta fitur copy-paste yang memungkinkan siswa menyalin jawaban temannya tanpa memahami makna dari jawaban tersebut [4]. Hasil pemeriksaan menggunakan Turnitin menyebutkan bahwa tingkat kemiripan karya ilmiah di Indonesia mengalami peningkatan dari 24,15% pada tahun 2021 menjadi 24,38% pada tahun 2022 [5]. Selain itu, laporan dari Buletin K-PIN mencatat bahwa lebih dari 60% siswa SMP dan SMA di Indonesia pernah melakukan plagiarisme [6]. Temuan tersebut menunjukkan bahwa plagiarisme telah menjadi persoalan yang serius dalam dunia pendidikan.

Berbagai penelitian telah menggunakan *Natural Language Processing* (NLP) untuk mendeteksi kemiripan teks. Salah satu pendekatan yang banyak digunakan adalah *Bidirectional Encoder Representations from Transformers* (BERT) karena mampu memahami makna kalimat berdasarkan dua arah (*bidirectional*) [7]. Seiring perkembangannya, *BERT* memiliki turunan yaitu *Sentence-BERT* (*SBERT*) yang mampu menghasilkan representasi kalimat (*sentence embedding*) [8]. Salah satu model yang banyak digunakan adalah *All-MiniLM-L12 v2*. Model *All-MiniLM-L12 v2* memiliki 12 lapisan (12 *layers*) untuk menangkap teks semantik, dengan ukuran model yang lebih ringan, namun tetap mampu merepresentasikan makna kalimat dengan baik [9].

Cosine similarity merupakan metode yang digunakan untuk menghitung tingkat kesamaan antar vektor teks dari dua teks. Nilai *cosine similarity* berada pada rentang 0 hingga 1, di mana nilai yang mendekati 1 menunjukkan kemiripan yang tinggi. Sedangkan untuk nilai 0 menunjukkan tingkat kemiripan yang rendah [10]. *Cosine similarity* banyak digunakan dalam penelitian *text similarity* karena mampu mengidentifikasi kemiripan makna antarteks secara efektif [11].

Pemilihan kombinasi *BERT* dan *cosine similarity* didasarkan pada kemampuan *BERT* dalam memahami kemiripan makna antar kalimat meskipun sudah dilakukan parafrase. Dalam konteks jawaban siswa, plagiarisme biasanya tidak hanya melalui *copy-paste*, tetapi juga melalui perubahan struktur kalimat maupun penggunaan sinonim [12]. Penelitian sebelumnya [13] menunjukkan bahwa pendekatan berbasis *BERT* memiliki korelasi yang lebih tinggi terhadap penilaian manusia dibandingkan metode tradisional seperti TF-IDF. *BERT* memperoleh nilai korelasi Pearson sebesar 0,8533, lebih tinggi dibandingkan TF-IDF yang memperoleh nilai 0,5497, sehingga dinilai lebih mampu menangkap hubungan semantik antarteks.

Meskipun berbagai penelitian telah menunjukkan bahwa kombinasi *BERT* dan *cosine similarity* mampu meningkatkan kinerja deteksi kemiripan teks, sebagian besar penelitian masih berfokus pada dokumen akademik atau *benchmark* dataset. Selain itu, penelitian terdahulu belum banyak menerapkan pendekatan pada jawaban esai siswa yang dikumpulkan melalui Google Form maupun melibatkan guru mata pelajaran dalam pelabelan manual (*ground truth*) untuk proses evaluasi. Oleh

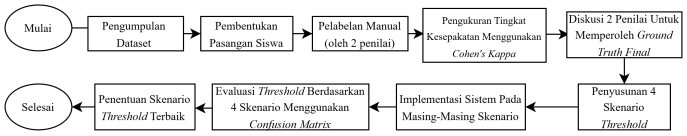
karena itu, penelitian ini mengembangkan sistem deteksi plagiarisme jawaban antar siswa berbasis Google Form menggunakan *Sentence-BERT all-MiniLM-L12-v2* dan metode *cosine similarity*.

Selanjutnya, empat skenario *threshold* dievaluasi menggunakan *confusion matrix* untuk memperoleh nilai *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil evaluasi tersebut akan digunakan sebagai dasar untuk menentukan *threshold* terbaik dalam mengklasifikasikan jawaban ke dalam kategori plagiarisme, parafrase, dan tidak plagiarisme.

II. METODOLOGI PENELITIAN

A. Alur Penelitian

Penelitian ini mengembangkan sistem untuk mendeteksi plagiarisme pada jawaban siswa yang dikumpulkan melalui Google Form, dengan menggunakan *Sentence-BERT* (*all-MiniLM-L12-v2*) dan metode *cosine similarity*. Penelitian bertujuan mengukur tingkat kemiripan semantik antarsiswa pada jawaban esai, dan menentukan batas ambang (*threshold*) yang menghasilkan kinerja terbaik dalam proses klasifikasi. Proses penelitian dilakukan secara bertahap, dimulai dengan: (1) mengumpulkan jawaban esai siswa dari Google Form dalam format CSV; (2) pembentukan pasangan siswa menggunakan metode perbandingan berpasangan (*pairwise comparison)*; (3) melakukan pelabelan secara manual oleh dua penilai yaitu peneliti dan guru mata pelajaran; (4) mengukur tingkat kesepakatan dua penilai menggunakan *Cohen's Kappa*; (5) mendiskusikan pasangan dengan label berbeda di antara kedua penilai untuk memperoleh *ground truth final*; (6) penyusunan empat skenario *threshold*; (7) mengimplementasi sistem pada setiap skenario *threshold* yang meliputi *text preprocessing*, pembentukan *sentence embedding* menggunakan *Sentence-BERT*, perhitungan nilai kemiripan menggunakan *cosine similarity*, dan klasifikasi jawaban berdasarkan *threshold*; (8) mengevaluasi *threshold* pada keempat skenario menggunakan *confusion matrix* untuk memperoleh nilai *precision*, *recall*, *F1-score*, dan *accuracy*; serta (9) menentukan skenario *threshold* terbaik berdasarkan hasil evaluasi. Alur penelitian secara keseluruhan ditunjukkan pada Gbr 1.



Gbr. 1 Diagram Alur Penelitian

B. Dataset

Dataset yang digunakan dalam penelitian ini adalah jawaban esai siswa kelas XI TKR 2 di SMKN 1 Baureno dengan mata pelajaran sasis, yang dikumpulkan melalui Google Form. Dataset tersebut mencakup jawaban dari 24 siswa untuk lima pertanyaan esai. Seluruh jawaban diekspor ke dalam format CSV untuk memudahkan pemrosesan data. Setiap jawaban kemudian dipasangkan menggunakan metode perbandingan berpasangan (*pairwise comparison*).

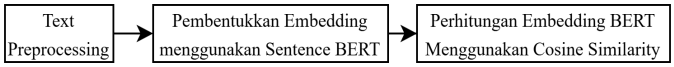
C. Pelabelan Manual dan Perhitungan Cohen's Kappa

Setelah proses pembentukan pasangan siswa, pelabelan dilakukan oleh dua penilai, yaitu peneliti dan guru mata pelajaran secara independen. Pelabelan ini melibatkan proses klasifikasi jawaban pada setiap nomor soal ke dalam tiga kategori, yaitu plagiarisme, parafrase, dan tidak plagiarisme.

Selanjutnya, hasil pelabelan dari kedua penilai dibandingkan menggunakan *Cohen's Kappa* untuk mengetahui tingkat kesepakatan. Data yang diberi kategori sama oleh kedua penilai langsung ditetapkan sebagai *ground truth*, sedangkan data yang memiliki kategori berbeda tidak langsung digunakan sebagai *ground truth*. Namun, data tersebut selanjutnya akan didiskusikan untuk meninjau kembali jawaban siswa dan mencapai kesepakatan mengenai kategori yang paling tepat. Setelah kesepakatan tercapai, label tersebut ditetapkan sebagai *ground truth final* yang menjadi acuan dalam proses evaluasi seluruh skenario *threshold*.

D. Implementasi Sistem

Implementasi sistem deteksi plagiarisme dilakukan melalui beberapa tahapan: *text preprocessing*, pembentukan *sentence embedding* menggunakan *Sentence-BERT* (*all-MiniLM-L12-v2*), perhitungan skor kemiripan menggunakan *cosine similarity*, dan klasifikasi berdasarkan *threshold*. Alur implementasi sistem ditunjukkan pada Gbr 2.



Gbr. 2 Diagram Alur Implementasi Sistem

1) Text Preprocessing

Pra-pemrosesan teks dilakukan untuk menormalisasi teks dan menghilangkan *noise*, sehingga menghasilkan data yang lebih konsisten serta meningkatkan kualitas representasi kalimat (*sentence embedding*) dan perhitungan tingkat kemiripan. Tahapan yang digunakan pada penelitian ini ditunjukkan pada Tabel I.

TABEL I
Tahapan Text Preprocessing

Tahapan	Deskripsi
Case Folding	Semua huruf kapital pada jawaban siswa akan diubah menjadi huruf kecil untuk menyamakan format penulisan, misalnya ”Roda” dan “roda”.
R*egular Expression* (*Regex*)	Membersihkan teks dari karakter nonalfanumerik yang tidak relevan, seperti simbol, karakter khusus, dan *whitespace* berlebih.
Filtering	Menghapus kata-kata yang tidak memiliki makna penting, seperti kata yang termasuk dalam *stopwords,* antara lain: ”yang”, “dan”, “di”, “itu”, “dengan”, “untuk”, “tidak”, “dari”, dst.
Tokenizing	Tahapan untuk memecah jawaban siswa menjadi unit-unit kata (*token*).

2) Ekstraksi Embedding dengan Sentence-BERT

Jawaban siswa yang telah melalui proses *text preprocessing* digunakan sebagai *input* pada *sentence-BERT* (*all-MiniLM-L12-v2)*. Model *all-MiniLM-L12-v2* terdiri atas 12 lapisan (12 layers) yang dirancang untuk menghasilkan representasi semantik kalimat, namun dengan komputasi yang lebih ringan apabila dibandingkan dengan model *BERT* berukuran besar [14]. Setiap jawaban siswa diubah menjadi representasi numerik berupa *sentence embedding* berdimensi 384 yang merepresentasikan makna dari kalimat.

3) Cosine Similarity

Representasi *sentence embedding* yang berdimensi 384 yang dihasilkan oleh model *all-MiniLM-L12-v2* digunakan sebagai *input* pada metode *cosine similarity*. *Cosine similarity* menghitung tingkat kemiripan antarvektor dan menghasilkan skor *similarity* pada rentang 0 hingga 1 [15]. Nilai yang mendekati 1 menunjukkan tingkat kemiripan semantik yang lebih tinggi, sedangkan nilai yang mendekati 0 menunjukkan tingkat kemiripan yang lebih rendah.

Perhitungan *cosine similarity* dilakukan dengan membandingkan dua vektor *sentence embedding*. Perhitungan diawali dengan menghitung nilai pembilang menggunakan dot (*dot product*), yaitu perkalian setiap elemen yang bersesuaian dari kedua vektor *embedding*. Selanjutnya, menghitung nilai penyebut yang berupa hasil perkalian norma dari masing-masing vektor *embedding*. Perbandingan antara nilai pembilang dan penyebut menghasilkan nilai *similarity* yang selanjutnya digunakan sebagai dasar klasifikasi jawaban berdasarkan *threshold*. Perhitungan *cosine similarity* ditunjukkan pada Persamaan (1)

$$\frac{\sum_i^n A_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}} \quad (1)$$

E. Penentuan Threshold

Penelitian ini menggunakan empat skenario *threshold* untuk menetapkan batas klasifikasi hasil skor perhitungan *cosine similarity* ke dalam kategori plagiarisme, parafrase, dan tidak plagiarisme. Setiap skenario menggunakan rentang nilai *similarity* yang berbeda, sehingga menghasilkan klasifikasi yang berbeda. Keempat skenario tersebut diterapkan pada sistem secara bergantian, dan hasil klasifikasinya akan dibandingkan dengan *ground truth final* untuk menentukan skenario mana yang memberikan performa terbaik.

TABEL II
Skenario *Threshold*

Skenario	Plagiarisme	Parafrase	Tidak Plagiarisme
1	≥ 90%	70% – 89%	< 70%
2	≥ 80%	60% – 79%	< 60%
3	≥ 70%	50% – 69%	< 50%
4	≥ 60%	40% – 59%	< 40%

F. Evaluasi Threshold

Evaluasi dilakukan untuk mengukur kinerja sistem dalam mendeteksi plagiarisme pada berbagai skenario *threshold*. Hasil klasifikasi dari masing-masing skenario dibandingkan dengan *ground truth final* menggunakan *confusion matrix* untuk memperoleh nilai *precision*, *recall*, *F1-score*, dan *accuracy* sebagai indikator performa sistem.

1.) Precision

Digunakan untuk mengukur tingkat ketepatan sistem dalam mengklasifikasikan suatu kategori. Nilai *precision* menunjukkan proporsi data yang diprediksi dengan benar dalam suatu kategori dibandingkan dengan seluruh data yang diprediksi masuk ke dalam kategori tersebut. Semakin tinggi nilai *precision*, maka akan semakin kecil kemungkinan sistem menghasilkan prediksi positif yang keliru. Rumus perhitungan *precision* ditunjukkan pada Persamaan (2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

2.) Recall

Digunakan untuk mengukur kemampuan sistem dalam mengidentifikasi seluruh data yang seharusnya termasuk dalam suatu kategori. Nilai *recall* menunjukkan proporsi data aktual yang berhasil diidentifikasi oleh sistem. Rumus perhitungan untuk *recall* ditunjukkan pada Persamaan (3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

3.) F1-score

Merupakan rata-rata harmonik dari *precision* dan *recall*, sehingga memberikan ukuran kinerja yang menyeimbangkan akurasi dan kemampuan sistem dalam mengidentifikasi data. Rumus perhitungan *F1-score* ditunjukkan pada Persamaan (4).

$$F1 - Score = \frac{2x\ Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

4.) Accuracy

Digunakan untuk mengukur tingkat ketepatan klasifikasi sistem secara keseluruhan berdasarkan proporsi jumlah prediksi yang benar terhadap seluruh data uji. Rumus perhitungan *accuracy* ditunjukkan pada Persamaan (5).

$$Akurasi = \frac{Jumlah\ Prediksi\ Benar}{Total\ Data\ Uji} \quad (5)$$

III. HASIL DAN PEMBAHASAN

A. Hasil Pembentukan Pasangan

Dataset penelitian terdiri atas jawaban esai dari 24 siswa kelas XI TKR 2 SMKN 1 Baureno yang memiliki lima pertanyaan dalam bentuk esai. Jawaban siswa tersebut kemudian diproses menggunakan metode *pairwise comparison*, sehingga menghasilkan 276 pasangan. Karena terdapat 5 pertanyaan esai, maka akan menghasilkan 1.380 data yang

selanjutnya akan digunakan dalam proses pelabelan manual dan klasifikasi ke dalam kategori plagiarisme, parafrase, dan tidak plagiarisme. Ringkasan hasil pembentukan pasangan ditunjukkan pada Tabel III.

TABEL III
Hasil Pembentukan Pasangan Jawaban

Jumlah siswa	24
Jumlah pasangan siswa	276
Jumlah soal esai	5
Total data yang diklasifikasikan	1.380

Berdasarkan Tabel III, proses pembentukan pasangan siswa menggunakan metode *pairwise comparison* menghasilkan 276 pasangan dari 24 siswa. Selanjutnya, setiap pasangan dibandingkan pada lima soal esai, sehingga diperoleh 276 x 5 = 1.380 data klasifikasi.

B. Hasil Pelabelan Manual

TABEL IV
Hasil Pelabelan Manual

Penilai	Plagiarisme	Parafrase	Tidak Plagiarisme	Total
Peneliti	615	465	300	1.380
Guru	614	512	254	1.380

Berdasarkan Tabel IV, distribusi label antara peneliti dan guru mata pelajaran menunjukkan hasil yang hampir serupa. Pada kategori plagiarisme, peneliti memberikan label sebanyak 615 data, sedangkan guru sebanyak 614 data. Perbedaan jumlah pelabelan lebih banyak ditemukan pada kategori parafrase dan tidak plagiarisme, yang menunjukkan adanya perbedaan interpretasi dalam mengkategorikan beberapa hasil perbandingan jawaban pada setiap nomor soal. Oleh karena itu, dilakukan pengukuran tingkat kesepakatan menggunakan *Cohen's Kappa* untuk mengetahui konsistensi hasil pelabelan kedua penilai.

TABEL V
Hasil Perbandingan Pelabelan Kedua Penilai

Jumlah Label Sama	Jumlah Label Berbeda
1.288	92

Berdasarkan Tabel V, sebanyak 1.288 data (93,33%) memperoleh kategori yang sama antara peneliti dan guru mata pelajaran, sedangkan 92 data (6,67%) memperoleh kategori yang berbeda. Perbedaan tersebut menunjukkan bahwa terdapat sejumlah pasangan jawaban yang masih memerlukan peninjauan lebih lanjut untuk memperoleh kategori yang sesuai.

C. Hasil Cohen's Kappa

TABEL VI
Hasil Pengukuran *Cohen's Kappa*

Observed Agreement (Po)	*Expected Agreement* (*Pe*)	*Cohen's kappa* (*K*)
0,9333	0,3633	0,8954

Berdasarkan Tabel VI, diperoleh nilai Observed Agreement (Po) sebesar 0,9333, yang menunjukkan bahwa 93,33% hasil pelabelan antara peneliti dan guru memiliki kesesuaian. Sementara itu, nilai Expected Agreement (Pe) sebesar 0,3633 menunjukkan tingkat kesepakatan yang mungkin terjadi secara kebetulan. Berdasarkan kedua nilai tersebut diperoleh Cohen's Kappa sebesar 0,8954, yang termasuk dalam kategori almost perfect agreement. Hasil ini menunjukkan bahwa tingkat kesepakatan kedua penilai sangat tinggi sehingga ground truth final yang dihasilkan layak digunakan sebagai acuan dalam proses evaluasi seluruh skenario *threshold*.

D. Hasil Diskusi Kedua Penilai

Diskusi dilakukan terhadap 92 data yang memperoleh kategori berbeda pada proses pelabelan manual. Melalui proses tersebut, peneliti dan guru mata pelajaran menyepakati kategori akhir untuk setiap data sehingga diperoleh ground truth final yang digunakan sebagai acuan dalam proses evaluasi seluruh skenario *threshold*. Distribusi kategori hasil diskusi ditunjukkan pada Tabel VII.

TABEL VII
Distribusi Kategori *Ground Truth Final*

Kategori	Jumlah
Plagiarisme	621
Parafrase	451
Tidak Plagiarisme	308

Berdasarkan Tabel VII, hasil diskusi menghasilkan 621 data pada kategori plagiarisme, 451 data pada kategori parafrase, dan 308 data pada kategori tidak plagiarisme. Distribusi kategori tersebut selanjutnya digunakan sebagai *ground truth final* dalam proses evaluasi seluruh skenario *threshold*.

E. Hasil Evaluasi Threshold

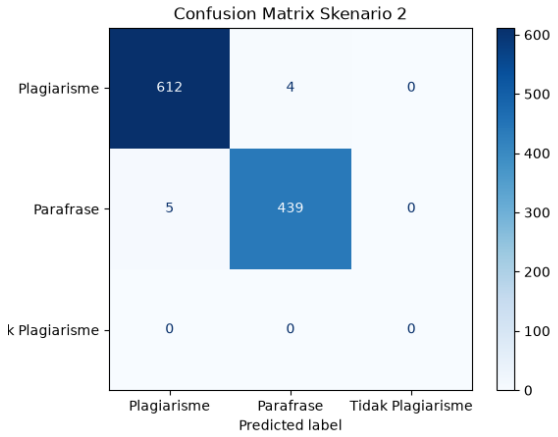
TABEL VIII
Hasil Evaluasi Skenario Threshold

Skenario	Akurasi
1	61,30%
2	94,42%
3	78,55%
4	62,89%

Berdasarkan Tabel VIII, setiap skenario *threshold* menghasilkan tingkat akurasi yang berbeda. Skenario 2 memperoleh akurasi tertinggi sebesar 94,42%, sehingga dipilih sebagai *threshold* terbaik pada penelitian ini. Hasil tersebut menunjukkan bahwa skenario 2 mampu memberikan

keseimbangan terbaik dalam membedakan kategori plagiarisme, parafrase, dan tidak plagiarisme.

Sementara itu, skenario 1 hanya memperoleh akurasi sebesar 61,30% karena penggunaan *threshold* yang terlalu tinggi menyebabkan banyak data yang seharusnya termasuk kategori plagiarisme diklasifikasikan ke dalam kategori parafrase atau tidak plagiarisme. Sebaliknya, skenario 4 memperoleh akurasi sebesar 62,89% akibat penggunaan *threshold* yang terlalu rendah sehingga banyak data diklasifikasikan sebagai plagiarisme. Adapun skenario 3 menghasilkan akurasi sebesar 78,55%, namun masih lebih rendah dibandingkan skenario 2 karena batas *threshold* yang digunakan belum mampu memisahkan kategori plagiarisme dan parafrase secara optimal.



Gbr. 3 Confusion Matrix Skenario 2

Berdasarkan Gbr. 3, sebagian besar data pada setiap kategori berhasil diklasifikasikan dengan benar oleh sistem. Hal ini ditunjukkan pada diagonal utama *confusion matrix* sebanyak 612 data kategori plagiarisme, 439 data pada kategori parafrase, dan 252 data pada kategori tidak plagiarisme. Namun, masih terdapat beberapa kesalahan klasifikasi yaitu sebanyak 53 data pada kategori tidak plagiarisme yang diprediksi sebagai parafrase. Secara umum, *confusion matrix* menunjukkan bahwa skenario 2 mampu memberikan hasil klasifikasi yang sesuai dengan *ground truth*.

TABEL IX
Hasil Evaluasi Skenario 2

Kategori	Precision	Recall	F1-Score
Plagiarisme	0,99	0,99	0,99
Parafrase	0,88	0,97	0,93
Tidak Plagiarisme	0,96	0,82	0,88

Berdasarkan Tabel IX, skenario 2 menunjukkan performa yang baik pada ketiga kategori klasifikasi. Kategori plagiarisme memperoleh nilai *precision*, *recall*, dan *F1-score* masing-masing sebesar 0,99, yang menunjukkan bahwa hampir seluruh data pada kategori tersebut berhasil diklasifikasikan dengan benar.

Pada kategori parafrase, diperoleh nilai *precision* sebesar 0,88, *recall* sebesar 0,97, dan *F1-score* sebesar 0,93. Hasil ini menunjukkan bahwa sistem mampu mengenali sebagian besar data parafrase dengan baik, meskipun masih terdapat beberapa data yang diklasifikasikan ke kategori lain. Adapun pada kategori tidak plagiarisme, nilai *precision* sebesar 0,96, *recall* sebesar 0,82, dan *F1-score* sebesar 0,88. Nilai *recall* yang lebih rendah pada kategori tidak plagiarisme menunjukkan bahwa masih terdapat beberapa data yang seharusnya masuk ke kategori tidak plagiarisme, namun sistem masih belum berhasil mengenalinya dan dimasukkan ke kategori yang lain. Secara keseluruhan, hasil evaluasi tersebut menunjukkan bahwa skenario 2 memberikan performa klasifikasi yang baik sehingga dipilih sebagai *threshold* terbaik pada penelitian ini.

TABEL X
Contoh Kasus Kesalahan Klasifikasi

Jawaban Siswa A	Jawaban Siswa B	Skor Similarity	Prediksi Sistem	*Ground Truth*
Dari komponen yang disebutkan, output shaft adalah yang paling krusial. Output shaft adalah jalur akhir yang menyalurkan seluruh tenaga yang sudah diatur oleh komponen lain langsung ke roda, menjadikannya 'jembatan' terakhir yang tak tergantikan.	Input shaft adalah komponen paling krusial	74,12%	Parafrase	Tidak Plagiarisme

Berdasarkan Tabel X, sistem mengklasifikasikan pasangan jawaban tersebut sebagai kategori parafrase dengan skor kemiripan 74,12%, sedangkan *ground truth* melabeli pasangan jawaban tersebut sebagai tidak plagiarisme. Kesalahan klasifikasi ini terjadi karena kedua jawaban membahas alasan yang sama, yaitu komponen yang paling krusial dalam mentransmisikan tenaga mesin ke roda, sehingga menghasilkan tingkat kemiripan semantik yang relatif tinggi. Padahal informasi kunci yang disampaikan dalam setiap jawaban berbeda. Siswa A memilih *output shaft,* sedangkan siswa B memilih *input shaft*.

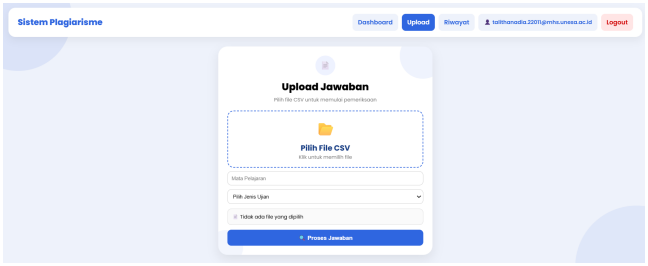
Kasus serupa tidak hanya ditemukan pada satu pasangan jawaban, tetapi juga ditemukan pada beberapa data lainnya, terutama pada soal yang memiliki pola alasan jawaban hampir sama. Hal ini menunjukkan bahwa *sentence-BERT* lebih sensitif terhadap kesamaan konteks kalimat dibandingkan perbedaan pada 9 utama jawaban.

F. Hasil Implementasi Sistem

Sistem deteksi plagiarisme diimplementasikan menggunakan *threshold* pada skenario dua yang dipilih sebagai *threshold* terbaik berdasarkan hasil evaluasi dengan akurasi 94,42%. *Threshold* tersebut digunakan sebagai dasar untuk mengklasifikasikan ke dalam tiga kategori, yaitu plagiarisme, parafrase, dan tidak plagiarisme. Hasil

implementasi sistem ditampilkan dalam beberapa halaman antarmuka, meliputi halaman *upload*, halaman *ranking* plagiarisme, halaman detail perbandingan, dan halaman riwayat pemeriksaan.

1.) *Halaman Upload*



Gbr. 4 Halaman Upload

Pada halaman *upload*, pengguna harus mengunggah jawaban siswa dalam format CSV, kemudian mengisi mata pelajaran dan memilih jenis ujian yaitu UH, UTS atau UAS. Setelah data diunggah, sistem akan memproses jawaban siswa dan menampilkan hasil deteksi plagiarisme.

2.) *Hasil Ranking Plagiarisme*

Gbr. 5 Halaman Ranking Plagiarisme

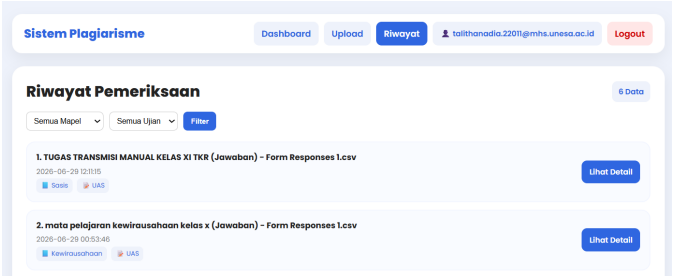
Sistem menampilkan hasil deteksi dalam bentuk *ranking* berdasarkan jumlah keterlibatan siswa yang melakukan plagiarisme. Apabila semakin banyak nomor soal yang terindikasi sebagai plagiarisme, maka akan semakin tinggi posisi siswa pada halaman *ranking*.

3.) *Halaman Detail Perbandingan*

Gbr. 6 Halaman Detail Perbandingan

Halaman detail perbandingan menampilkan daftar pasangan siswa yang terindikasi melakukan plagiarisme dengan siswa yang dipilih pada halaman *ranking*. Pengguna juga dapat memfilter hasil detail perbandingan berdasarkan nomor soal, sehingga sistem hanya akan menampilkan daftar pasangan siswa yang terindikasi plagiarisme pada nomor soal yang dipilih.

4.) *Halaman Riwayat Pemeriksaan*



Gbr. 7 Halaman Riwayat Pemeriksaan

Halaman riwayat menampilkan nama file, mata pelajaran, dan jenis ujian yang sebelumnya di*input*kan oleh pengguna pada halaman upload. Halaman ini berfungsi untuk mempermudah pengguna dalam melihat kembali hasil deteksi tanpa harus mengunggah data jawaban CSV lagi.

IV. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem deteksi plagiarisme jawaban antarsiswa berbasis Google Form menggunakan *Sentence-BERT* (*all-MiniLM-L12-v2*) dan metode *cosine similarity*. Hasil pelabelan manual oleh peneliti dan guru mata pelajaran menunjukkan nilai Cohen's Kappa sebesar 0,8954, yang termasuk kategori *almost perfect agreement*, sehingga *ground truth* yang dihasilkan layak digunakan sebagai acuan evaluasi. Berdasarkan pengujian empat skenario *threshold*, skenario 2 dengan batas ≥80% untuk plagiarisme, 60%–79% untuk parafrase, dan <60% untuk tidak plagiarisme memberikan performa terbaik dengan nilai akurasi sebesar 94,42%.

UCAPAN TERIMA KASIH

Penulis mengucapkan puji dan syukur kepada Allah SWT atas rahmat, karunia dan pertolongan-Nya sehingga penelitian ini dapat diselesaikan dengan baik. Penulis juga menyampaikan terima kasih kepada dosen pembimbing atas bimbingan, arahan, masukan, serta motivasi yang telah diberikan selama proses penyusunan dan penyelesaian penelitian ini. Ucapan terima kasih juga disampaikan kepada SMKN 1 Baureno atas izin, dukungan, serta membantu proses pengumpulan data dan pelabelan manual. Selain itu, penulis juga mengucapkan terima kasih kepada keluarga, teman-teman, serta seluruh pihak yang telah memberikan dukungan, bantuan, serta semangat selama proses penelitian hingga artikel ini dapat diselesaikan. Semoga segala bantuan, dukungan, dan kebaikan yang telah diberikan mendapatkan balasan terbaik dari Allah SWT.

REFERENSI

[1] A. Awaluddin, F. Ramadan, F. A. N. Charty, and and M. Firmansyah, “Peran Pengembangan dan

Pemanfaatan Teknologi Pendidikan dan Pembelajaran Dalam Meningkatkan Kualitas Mengajar," *Jurnal Pendidikan Teknologi Informasi*, vol. 2 No 2, pp. 48–59, Jul. 2021, doi: https://e-journal.unimudasorong.ac.id/index.php/jurnalpetisi/article/view/801/214.

[2] A. A. Widati *et al.*, "Pemanfaatan Teknologi Digital sebagai Media Pembelajaran untuk Guru SMA-SMP Muhammadiyah Gresik," *Bubungan Tinggi: Jurnal Pengabdian Masyarakat*, vol. 5, no. 2, p. 679, May 2023, doi: 10.20527/btjpm.v5i2.6873.

[3] M. A. Pratiwi and N. Aisya, "Fenomena plagiarisme akademik di era digital," *Publishing Letters*, vol. 1, no. 2, pp. 16–33, Jul. 2021, doi: 10.48078/publetters.v1i2.23.

[4] Anis Humaidi, Najihatul Fadhliyah, and Dina Ameliana, "DAMPAK TEKNOLOGI TERHADAP MARAKNYA PLAGIARISME DIKALANGAN MAHASISWA (Studi Kasus di UIN Syekh Wasil Kediri)," *Studi Pendidikan Agama Islam*, vol. 8, no. 1, pp. 384–399, Mar. 2026, doi: 10.54437/ilmuna.

[5] A. A. Prakoso, N. Fadilah, and A. Yahya Maulana, "Analisis Tingkat Plagiarisme Terhadap Karya Ilmiah Jurnal Mahasiswa Perpustakaan dan Sains Informasi Universitas YARSI Jakarta Tahun 2021-2022 dengan Software Turnitin," *Jurnal Pustaka Ilmiah*, vol. 10, no. 1, pp. 115–128, Jun. 2024, doi: 10.20961/jpi.v10i1.84924.

[6] D. Syukriah, "Plagiarisme Dalam Dunia Pendidikan Di Indonesia," *Buletin KPIN*, vol. 8, no. 17, Aug. 2022. [Online]. Available: www.nasional.tempo.com.

[7] A. Febrianti Azzaroh, A. Surya Editya, and A. Khoir Al-Haq, "Sistem Penilaian Esai Otomatis Berbasis Web Menggunakan Model BERT dan Levenshtein Distance," *Jurnal Pustaka AI (Pusat Akses Kajian Teknologi Artificial Intelligence)*, vol. 5, no. 2, pp. 142–148, Aug. 2025, doi: 10.55382/jurnalpustakaai.v5i2.1022.

[8] J. Opitz and A. Frank, "SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features," Oct. 2022. doi: 10.18653/v1/2022.aacl-main.48.

[9] S. Srivastava, A. Devi, N. Kriti, S. Uttrani, K. Pudari, and V. Dutt, "Evaluating Accuracy and Speed in Similarity Models for Medical Data," 2025.

[10] T. H. Syah, S. Syah, and N. Nurmalasari, "Minat Masyarakat Terhadap Karakteristik Informasi Tekstual dari Kementerian Lingkungan Hidup dan Kehutanan di Media Sosial," *Jurnal Komunika: Jurnal Komunikasi, Media dan Informatika*, vol. 10, no. 2, pp. 62–72, Dec. 2021, doi: 10.31504/komunika.v10i2.4442.

[11] R. Harjo Utomo, G. Susrama, M. Diyasa, U. Pembangunan, N. " Veteran, and J. Timur, "MOVIEMU : SISTEM REKOMENDASI FILM MENGGUNAKAN ALGORITMA COSINE SIMILARITY," *Pengabdian Masyarakat SENSASI / Articles*, vol. 4, no. 2, pp. 22–32, 2024, doi: https://doi.org/10.33005/sensasi.v4i2.83.

[12] M. A. Pratiwi and N. Aisya, "Fenomena plagiarisme akademik di era digital," *Publishing Letters*, vol. 1, no. 2, pp. 16–33, Jul. 2021, doi: 10.48078/publetters.v1i2.23.

[13] Maulidya Prastita Syah, Ajeng Puspa Wardani, Mohammad Idhom, and Trimono, "Perbandingan Representasi Teks Tf-Idf Dan Bert Terhadap Akurasi Cosine Similarity Dalam Penilaian Otomatis Jawaban Berbasis Teks," *Data Sciences Indonesia (DSI)*, vol. 5, no. 1, pp. 47–59, Jul. 2025, doi: 10.47709/dsi.v5i1.6021.

[14] Z. Xian and T. Zhuhai, "4 An Effective Approach to Embedding Source Code by Combining Large Language and Sentence Embedding Models," *Journal of the ACM*, vol. 37, Jun. 2025, doi: https://doi.org/10.48550/arXiv.2409.14644.

[15] S. Agustian, "Sistem Qur'an Retrieval Terjemahan Bahasa Indonesia Berbasis Web dengan Reorganisasi Korpus", doi: 10.13140/2.1.5168.6083.