

SISTEM SCORING TUGAS AKHIR BERDASARKAN KEMIRIPAN SEMANTIK JUDUL DAN DESKRIPSI MENGGUNAKAN INDOBERT

Ahmad Sihabillah¹, Moh. Furqan¹, Fathorazi Nur Fajri^{*1}

¹Program Studi Teknik Informatika, Fakultas Teknik, Universitas Nurul Jadid, Probolinggo, Indonesia

Penulis korespondensi: Fathorazi Nur Fajri (fathorazi@unuja.ac.id)

Evaluasi awal usulan tugas akhir umumnya masih dilakukan secara manual sehingga memerlukan waktu yang lama dan berpotensi menghasilkan penilaian yang kurang konsisten. Penelitian ini bertujuan mengembangkan sistem berbasis kemiripan semantik menggunakan IndoBERT yang dioptimalkan dengan pendekatan SimCSE untuk membantu mengidentifikasi tingkat kemiripan dan keunikan usulan tugas akhir. Model dilatih menggunakan tiga jenis pasangan positif dan dievaluasi pada skenario judul, deskripsi, serta gabungan keduanya menggunakan metrik Precision@5, Recall@5, dan F1-Score@5. Hasil penelitian menunjukkan bahwa optimasi model meningkatkan selisih rata-rata nilai cosine similarity antara pasangan positif dan negatif dari 0,040 menjadi 0,5412. Skenario deskripsi dan gabungan menghasilkan F1-Score@5 tertinggi sebesar 0,71 pada threshold 0,6. Model kemudian berhasil diimplementasikan pada aplikasi berbasis web dan memperoleh tingkat penerimaan pengguna sebesar 84,40%. Hasil tersebut menunjukkan bahwa pendekatan SimCSE efektif dalam mendukung evaluasi awal usulan tugas akhir secara lebih cepat dan objektif.

Kata Kunci — IndoBERT, SimCSE, kemiripan semantik, cosine similarity, tugas akhir.

I. PENDAHULUAN

Pemilihan judul tugas akhir merupakan tahap awal yang menentukan arah dan kualitas penelitian mahasiswa [1]. Namun, proses ini masih menjadi tantangan karena mahasiswa sering mengalami kesulitan menemukan topik yang orisinal sekaligus relevan dengan perkembangan penelitian sebelumnya. Penelitian menunjukkan bahwa 14 dari 46 mahasiswa mengalami kebingungan dalam menentukan topik penelitian [2]. Selain itu, 54,58% dari 71 mahasiswa semester akhir mengalami kendala dalam menemukan referensi yang sesuai [3]. Di sisi lain, jumlah publikasi ilmiah terus meningkat sekitar 4,10% per tahun sehingga proses penelusuran literatur menjadi semakin kompleks [4]. Kondisi tersebut menyebabkan proses evaluasi usulan tugas akhir menjadi semakin sulit apabila

hanya dilakukan secara manual. Oleh karena itu, diperlukan suatu pendekatan yang mampu mengukur tingkat kemiripan semantik antara usulan penelitian dengan penelitian yang telah ada secara lebih objektif dan efisien. Untuk menjawab kebutuhan tersebut, berbagai penelitian telah mengembangkan metode pengukuran kemiripan teks dengan pendekatan yang berbeda-beda. Berbagai penelitian telah mengembangkan sistem pengukuran kemiripan teks menggunakan TF-IDF maupun model *embedding*. Pendekatan berbasis TF-IDF masih kurang mampu menangkap hubungan semantik ketika terdapat variasi struktur kalimat [5], sedangkan penelitian berbasis transformer sebagian besar masih menggunakan korpus berbahasa Inggris atau dataset berbahasa Indonesia yang terbatas [6], [7]. Salah satu model transformer yang banyak digunakan untuk bahasa Indonesia adalah IndoBERT. Meskipun IndoBERT memiliki performa yang baik pada berbagai tugas NLP berbahasa Indonesia [8], representasi *embedding* yang dihasilkan tanpa *fine-tuning* belum dioptimalkan untuk tugas *semantic textual similarity* [9]. Arsitektur BERT dirancang sebagai *encoder* kontekstual umum [10], sedangkan pembentukan *embedding* kalimat untuk pencarian kemiripan memerlukan strategi representasi khusus agar dapat dibandingkan secara efisien menggunakan *cosine similarity* [11]. Analisis geometri *embedding* BERT juga menunjukkan bahwa representasi kontekstual tidak selalu bersifat isotropik, sehingga optimasi representasi kalimat relevan untuk tugas kemiripan semantik [12].

Di sisi lain, pendekatan SimCSE telah digunakan untuk meningkatkan kualitas representasi kalimat melalui *contrastive learning* [13]. Namun, sebagian besar penelitian sebelumnya hanya membentuk pasangan teks yang memiliki makna serupa serupa (*positive pair*) dari kalimat yang sama. Padahal, pada usulan tugas akhir, judul dan deskripsi merepresentasikan satu ide penelitian yang memiliki hubungan semantik kuat, tetapi pemanfaatannya sebagai *positive pair* masih sangat terbatas.

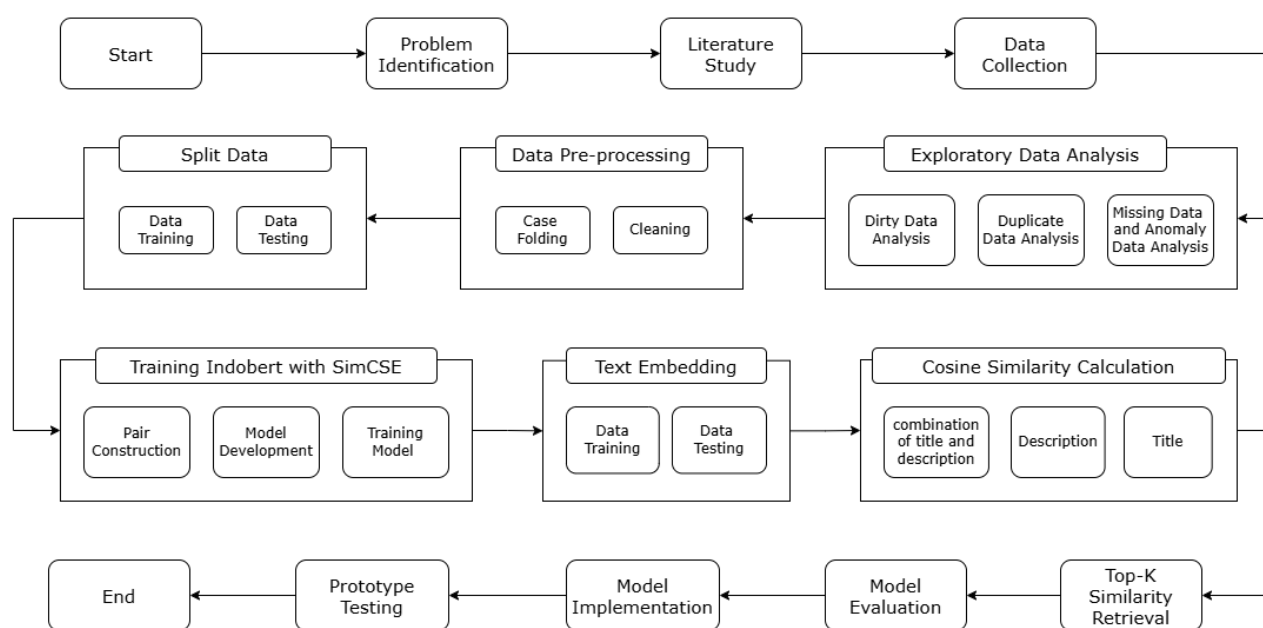
Berdasarkan kondisi tersebut, penelitian ini

mengusulkan pengoptimalan model IndoBERT menggunakan *unsupervised* SimCSE dengan memanfaatkan tiga jenis *positive pair*, yaitu *Self Pair* (pasangan dari kalimat yang sama dengan *noise dropout* berbeda), *Weakly Supervised Pair* (pasangan judul–deskripsi dari satu usulan tugas akhir yang sama), dan *Augmented Pair* (pasangan hasil augmentasi teks). Ketiga jenis pasangan tersebut dimanfaatkan sebagai sinyal kemiripan saat *fine-tuning* untuk menghasilkan representasi *embedding* yang lebih diskriminatif. Model yang dihasilkan selanjutnya digunakan untuk membandingkan judul dan deskripsi usulan penelitian yang diajukan terhadap basis data tugas akhir yang telah ada, sehingga tingkat kemiripannya dapat diidentifikasi secara lebih objektif sebagai indikator awal.

II. METODE PENELITIAN

Penelitian ini menerapkan pendekatan kuantitatif melalui eksperimen komputasional untuk mengembangkan model pengukuran kemiripan semantik pada usulan tugas akhir mahasiswa. Eksperimen dilakukan dengan mengoptimalkan model IndoBERT menggunakan pendekatan SimCSE, kemudian menguji kemampuan model dalam menemukan dokumen yang memiliki kedekatan semantik terhadap teks masukan. Performa model dinilai menggunakan metrik Precision@5, Recall@5, dan F1-Score@5.

Secara umum, penelitian dilaksanakan melalui beberapa tahapan yang meliputi identifikasi masalah, studi literatur, pengumpulan data sampai pengujian prototype. Rangkaian tahapan penelitian ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

A. Pengumpulan Data

Penelitian memanfaatkan dataset berupa judul dan deskripsi tugas akhir mahasiswa Fakultas Teknik Universitas Nurul Jadid periode 2012–2023. Dataset diperoleh dari repositori tugas akhir fakultas dan berisi 1.670 dokumen dalam format Microsoft Excel (.xlsx). Seluruh data digunakan sebagai sumber utama untuk proses pelatihan maupun evaluasi model.

B. Analisis Data Eksploratif

Sebelum memasuki tahap pemodelan, dilakukan *Exploratory Data Analysis* (EDA) untuk memperoleh gambaran mengenai kualitas dataset. Analisis difokuskan pada identifikasi data kosong, data duplikat, serta data yang tidak memenuhi kriteria sehingga dapat diketahui bentuk penanganan yang diperlukan sebelum proses pelatihan model dilakukan.

C. Pra-pemrosesan Data

Tahap pra-pemrosesan bertujuan menghasilkan data yang lebih konsisten dan siap digunakan pada proses *fine-tuning*. Tahapan ini mencakup *cleaning* data untuk menghilangkan data yang tidak valid, karakter yang tidak

diperlukan, serta data duplikat, kemudian dilanjutkan dengan *case folding* agar seluruh teks memiliki format penulisan yang seragam. Proses *stopword removal* dan *stemming* tidak diterapkan karena IndoBERT dirancang untuk memanfaatkan konteks kalimat secara utuh.

D. Pembagian Data

Dataset yang telah dilakukan pra-pemrosesan selanjutnya dibagi menjadi data latih dan data uji menggunakan rasio 90:10. Data latih digunakan dalam proses *fine-tuning* sekaligus sebagai basis referensi pada pencarian dokumen serupa, sedangkan data uji dimanfaatkan untuk mengevaluasi kemampuan model dalam menemukan dokumen yang relevan.

E. Pelatihan IndoBERT dengan SimCSE

Penelitian ini menggunakan model *pretrained* IndoBERT (*indobenchmark/indobert-base-p1*) yang dioptimalkan melalui pendekatan SimCSE menggunakan *framework* SentenceTransformers. Sebelum proses pelatihan, data latih dibentuk menjadi tiga variasi pasangan positif (*positive pair*), yaitu *Self Pair*, *Weakly Supervised Pair* yang berasal dari pasangan judul dan deskripsi pada dokumen yang sama, serta *Augmented Pair*

yang dihasilkan melalui augmentasi teks. Ketiga jenis pasangan tersebut digunakan sebagai masukan pada proses *fine-tuning* untuk mempelajari representasi semantik yang lebih diskriminatif. Selama pelatihan, pasangan negatif tidak dibentuk secara eksplisit, melainkan diperoleh secara otomatis dari pasangan positif lain dalam satu *batch* menggunakan fungsi *MultipleNegativesRankingLoss* sehingga proses pelatihan menjadi lebih efisien dan efektif untuk tugas pencarian informasi (*information retrieval*) serta pembelajaran representasi kalimat (*sentence embedding*)[14]. Proses pelatihan dilakukan menggunakan *batch size* 16, *epoch* 5, *learning rate* 3×10^{-5} , dan *maximum sequence length* 512 hingga diperoleh model SimCSE-IndoBERT yang digunakan pada tahap berikutnya. Pendekatan dense *retrieval* berbasis *embedding* telah terbukti efektif untuk pencarian semantik [15], sedangkan benchmark BEIR menekankan pentingnya evaluasi model *retrieval* pada beragam skenario dan domain [16].

F. Pembentukan Representasi Teks

Model hasil *fine-tuning* digunakan untuk mengubah setiap dokumen menjadi representasi vektor (*embedding*). Proses pembentukan *embedding* dilakukan pada tiga skenario, yaitu menggunakan judul, deskripsi, serta gabungan judul dan deskripsi. *Embedding* data latih disimpan sebagai *pre-computed embedding* yang berfungsi sebagai basis referensi, sedangkan *embedding* data uji digunakan untuk menghitung nilai *cosine similarity* terhadap seluruh *embedding* data latih. Hasil perhitungan tersebut selanjutnya digunakan pada tahap evaluasi model.

G. Perhitungan Kemiripan Cosine

Kemiripan semantik antara *embedding* data uji dan data latih dihitung menggunakan *cosine similarity*, yaitu ukuran kemiripan yang digunakan untuk membandingkan representasi vektor dari dua teks berdasarkan orientasi vektornya [17]. Nilai *cosine similarity* berada pada rentang -1 hingga 1 , dengan nilai yang semakin mendekati 1 menunjukkan bahwa kedua vektor memiliki arah yang semakin serupa [18]. Dalam penelitian ini, setiap *embedding* data uji dibandingkan dengan seluruh *embedding* data latih untuk memperoleh skor kemiripan, kemudian digunakan untuk menentukan dokumen dengan tingkat kemiripan semantik tertinggi.

H. Pengambilan Dokumen Termirip

Setelah seluruh nilai *cosine similarity* diperoleh, dokumen data latih diurutkan berdasarkan skor kemiripan dari yang tertinggi hingga terendah. Lima dokumen dengan skor tertinggi dipilih sebagai hasil *Top-5 Similarity Retrieval* yang merepresentasikan dokumen dengan kedekatan semantik paling tinggi terhadap dokumen uji. Hasil *retrieval* tersebut kemudian digunakan pada tahap evaluasi model untuk mengukur performa model yang diusulkan.

I. Evaluasi Model

Tahap evaluasi dilakukan untuk mengukur performa model dalam menghasilkan rekomendasi dokumen tugas akhir yang relevan. Pengujian dilakukan pada tiga skenario, yaitu berdasarkan judul, deskripsi, serta

gabungan judul dan deskripsi. Setiap skenario dievaluasi menggunakan metrik *Precision@5*, *Recall@5*, dan *F1-Score@5* pada variasi *threshold cosine similarity* 0,2 hingga 1,0 untuk menentukan konfigurasi yang memberikan performa terbaik.

J. Implementasi Model

Model hasil *fine-tuning* diimplementasikan ke dalam aplikasi berbasis web menggunakan framework *Flask*. Sistem menyediakan tiga mode pencarian, yaitu judul, deskripsi, serta gabungan judul dan deskripsi. *Embedding* data latih disimpan sebagai *pre-computed embedding* untuk mempercepat proses pencarian. Sistem kemudian menghitung *cosine similarity*, menampilkan lima dokumen termirip, serta menghasilkan skor keunikan berdasarkan nilai kemiripan tertinggi.

K. Pengujian Prototype

Prototype diuji menggunakan dua pendekatan, yaitu *Black Box Testing* dan *User Testing*. *Black Box Testing* dilakukan untuk memvalidasi fungsionalitas sistem berdasarkan input dan output tanpa memperhatikan implementasi kode program[19], sedangkan *User Testing* melibatkan 10 mahasiswa Fakultas Teknik sebagai responden untuk mengevaluasi kemudahan penggunaan sistem menggunakan kuesioner skala Likert.

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data (Data Collection)

Dataset yang digunakan pada penelitian ini diperoleh dari repositori tugas akhir Fakultas Teknik Universitas Nurul Jadid periode 2012–2023. Dataset terdiri atas 1.670 dokumen dalam format Microsoft Excel (.xlsx) yang memuat atribut *id_hasil_pengajuan_ujian*, *judul_ta*, *deskripsi*, dan *tahun*. Atribut *judul_ta* dan *deskripsi* digunakan sebagai sumber informasi utama dalam pembentukan pasangan teks dan representasi *embedding*, sedangkan *id_hasil_pengajuan_ujian* berfungsi sebagai identitas unik setiap dokumen dan *tahun* digunakan sebagai informasi pendukung. Cuplikan dataset yang digunakan pada penelitian ditunjukkan pada Tabel 1.

Tabel 1. Contoh Dataset

ID	Judul-Ta	Deskripsi	Tahun
1	Aplikasi perizinan santri di pondok pesantren Raudlatul Fatah di desa Maron berbasis Web	Dari penelitian di pondok pesantren Raudlatul Fatah berdasarkan data ... jika sukses akan input keterangan dari pengasuh, pengurus kamar, ketua kamar, dan keamanan.	2014
...	...	...	...
1670	Prediksi Keterlambatan Pembayaran Spp Menggunakan Metode Naive Bayes Di Smk Mambaul Ulum	NaN	2020

B. Analisis data eksploratif (*Exploratory Data Analysis*)

Analisis data eksploratif (*Exploratory Data Analysis* atau EDA) dilakukan untuk mengevaluasi kualitas dataset sebelum memasuki tahap *preprocessing*. Pemeriksaan ini difokuskan pada atribut judul_ta dan deskripsi. Adapun ringkasan hasil EDA disajikan pada Tabel 2.

Tabel 2. Hasil *Exploratory Data Analysis*

Jenis Pemeriksaan	Jumlah
Nilai NaN pada deskripsi	49
Teks pendek pada deskripsi	3
Duplikat judul	2
Duplikat deskripsi	7
Data kotor pada judul	19
Data kotor pada deskripsi	174
Data kotor pada kedua kolom	28

Berdasarkan Tabel 2, permasalahan kualitas data lebih banyak ditemukan pada kolom deskripsi dibandingkan judul_ta. Kolom deskripsi masih mengandung 49 nilai NaN, 3 teks pendek, dan 7 data duplikat, sedangkan kolom judul_ta hanya ditemukan 2 data duplikat tanpa adanya nilai NaN maupun teks pendek. Selain itu, pemeriksaan karakter tidak valid menunjukkan terdapat 221 data yang memerlukan pembersihan, dengan mayoritas ditemukan pada atribut deskripsi.

Temuan tersebut menunjukkan bahwa kualitas dataset perlu ditingkatkan sebelum digunakan pada proses pelatihan model. Oleh karena itu, seluruh permasalahan yang teridentifikasi ditangani pada tahap *pra-pemrosesan* untuk menghasilkan dataset yang lebih bersih.

C. Pra-pemrosesan Data (*Preprocessing Data*)

Tahap *pra-pemrosesan* dilakukan terhadap dataset hasil *Exploratory Data Analysis* (EDA) untuk meningkatkan kualitas dan konsistensi data sebelum digunakan pada proses *fine-tuning*. Tahapan yang diterapkan meliputi *cleaning* dan *case folding*.

Tabel 3. Contoh Hasil Pra-pemrosesan Data

Kolom	Sebelum <i>Preprocessing</i>	Setelah <i>Preprocessing</i>
judul_ta	Analisis dan Perancangan Sistem Akademik Berbasis WEB Pada Diniyah An-Nafi'iyah Wilayah Az-Zainiyah Pondok Pesantren Nurul Jadid	nalisis dan perancangan sistem akademik berbasis web pada diniyah an-nafi'iyah wilayah az-zainiyah pondok pesantren nurul jadid
deskripsi	Dewasa ini teknologi telah berkembang dengan pesat ... smart TV akan berada pada posisi ...	dewasa ini teknologi telah berkembang dengan pesat ... smart tv akan berada pada posisi on ...

Proses *cleaning* mencakup penghapusan data yang tidak valid, teks dengan panjang kurang dari tiga kata, serta data duplikat. Selain itu, dilakukan pembersihan karakter mojibake, penghapusan karakter yang tidak diperlukan, dan normalisasi spasi sehingga diperoleh representasi teks yang lebih bersih. Selanjutnya, *case folding* diterapkan dengan mengubah seluruh huruf menjadi huruf kecil (*lowercase*) agar penulisan setiap dokumen memiliki format yang seragam. Contoh hasil

preprocessing pada atribut judul_ta dan deskripsi ditunjukkan pada Tabel 3.

Setelah seluruh tahapan *preprocessing* selesai dilakukan, dataset divalidasi kembali menggunakan prosedur *Exploratory Data Analysis* (EDA) untuk memastikan bahwa seluruh permasalahan yang ditemukan pada tahap sebelumnya telah ditangani. Ringkasan hasil validasi setelah *preprocessing* disajikan pada Tabel 4.

Tabel 4. Hasil Validasi Dataset Setelah *Preprocessing*

Jenis Pemeriksaan	judul_ta	deskripsi
Nilai NaN	0	0
String Kosong	0	0
Teks pendek	0	0
Data Duplikat	1	4
Data Kotor (Eksklusif)	0	0
Data Kotor (Kedua Kolom)	0	0
Total Data Kotor	0	0

Berdasarkan Tabel 4, hasil validasi menunjukkan bahwa tidak ditemukan lagi nilai NaN, string kosong, teks pendek, maupun karakter yang tidak valid. Namun, proses normalisasi teks menyebabkan munculnya 1 data duplikat pada atribut judul_ta dan 4 data duplikat pada atribut deskripsi. Duplikasi tersebut muncul setelah proses *cleaning* dan *case folding* karena beberapa dokumen yang sebelumnya memiliki perbedaan karakter atau penggunaan huruf menjadi identik. Kelima data duplikat tersebut kemudian dihapus sehingga jumlah dataset berkurang dari 1.611 menjadi 1.606 dokumen. Dataset akhir yang telah dinyatakan bersih selanjutnya digunakan pada tahap pembagian data dan proses *fine-tuning* model IndoBERT menggunakan pendekatan SimCSE.

D. Pembagian Data (*Split Data*)

Proses pembagian dataset menghasilkan 1.445 data latih dan 161 data uji dari total 1.606 dokumen hasil *preprocessing*. Pembagian data dilakukan secara acak dengan proporsi 90% untuk data latih dan 10% untuk data uji agar model dapat dipelajari menggunakan sebagian besar data, sementara performanya tetap dapat dievaluasi pada data yang belum pernah dilihat sebelumnya. Ringkasan hasil pembagian data disajikan pada Tabel 5.

Tabel 5. Hasil Pembagian Data

Kelompok Data	Proporsi	Jumlah Data
Training	90%	1.445
Testing	10%	161
Total	100%	1.606

Hasil pembagian tersebut menunjukkan bahwa sebagian besar data dialokasikan sebagai data latih untuk mendukung proses *fine-tuning* model, sedangkan sebagian lainnya digunakan sebagai data uji untuk mengevaluasi kemampuan model pada data yang tidak terlibat selama proses pelatihan.

E. Pelatihan IndoBERT dengan SimCSE

Pelatihan model menghasilkan model SimCSE-IndoBERT yang selanjutnya digunakan pada tahap pembentukan representasi teks. Proses *fine-tuning*

memanfaatkan tiga jenis *positive pair*, yaitu *Self Pair*, *Weakly Supervised Pair*, dan *Augmented Pair*. Dari 1.445 data latih, proses pembentukan pasangan menghasilkan total 7.225 *positive pair* yang digunakan selama pelatihan. Ringkasan jumlah pasangan disajikan pada Tabel 6.

Tabel 6. Jumlah Pasangan Positif

Jenis Pasangan Positif	Jumlah Pasangan
<i>Self Pair</i>	2890
<i>Weakly Supervised Pair</i>	1445
<i>Augmented Pair</i>	2890

Ketiga jenis *positive pair* tersebut memiliki karakteristik yang berbeda. *Self Pair* mempertahankan teks asli sebagai pasangan positif, *Weakly Supervised Pair* memanfaatkan hubungan semantik antara judul dan deskripsi dari dokumen yang sama, sedangkan *Augmented Pair* dibentuk dari hasil augmentasi menggunakan teknik *random swap*. Seperti pada Tabel 7.

Tabel 7. Contoh Pasangan Positif

Kategori	Kandidat	Pasangan Positif
<i>Self Pair</i>	sistem e-complaint mahasiswa dosen dan karyawan universitas nurul jadid berbasis android	sistem e-complaint mahasiswa dosen dan karyawan universitas nurul jadid berbasis android
	sistem e-complaint mahasiswa dosen dan karyawan universitas nurul jadid berbasis android	sistem e-complaint ini bertujuan untuk mempermudah proses pengaduan
<i>Weakly Supervised Pair</i>	sistem e-complaint mahasiswa dosen dan karyawan universitas nurul jadid berbasis android	...
	sistem e-complaint mahasiswa dosen dan karyawan universitas nurul jadid berbasis android	disampaikan sesuai dengan hak akses masing-masing universitas jadid mahasiswa dosen dan karyawan sistem nurul e-complaint berbasis android
<i>Augmented Pair</i>	sistem e-complaint mahasiswa dosen dan karyawan universitas nurul jadid berbasis android	sistem e-complaint ini bertujuan untuk mempermudah proses pengaduan

Setelah seluruh pasangan berhasil dibentuk, proses *fine-tuning* menghasilkan model SimCSE-IndoBERT yang digunakan pada tahapan berikutnya. Nilai *training loss* selama proses pelatihan disajikan pada Tabel 8.

Tabel 8. Training Loss

Step	Training
500	0,170421
1000	0,031347

Berdasarkan Tabel 8, nilai *training loss* mengalami penurunan dari 0,170421 pada langkah ke-500 menjadi 0,031347 pada langkah ke-1000. Penurunan tersebut menunjukkan bahwa model semakin mampu mempelajari hubungan semantik antar teks selama proses *fine-tuning*. Selain itu, penurunan *training loss* yang konsisten mengindikasikan bahwa proses optimasi berlangsung dengan baik sehingga model mencapai konvergensi tanpa menunjukkan indikasi ketidakstabilan selama pelatihan.

Untuk mengevaluasi pengaruh *fine-tuning* terhadap kualitas representasi semantik, dilakukan perbandingan rata-rata nilai *cosine similarity* antara model IndoBERT Base dan SimCSE-IndoBERT menggunakan pasangan teks positif dan pasangan teks negatif. Hasil perbandingan disajikan pada Tabel 9.

Tabel 9. Perbandingan Nilai Rata-rata *Cosine Similarity* Pada Model IndoBERT Base dan SimCSE-IndoBERT

Model	Rata-rata Pasangan positif	Rata-rata Pasangan Negatif
IndoBERT Base	0,5593	0,5995
SimCSE-IndoBERT	0,6767	0,1355

Berdasarkan Tabel 9, model IndoBERT Base menghasilkan rata-rata *cosine similarity* pasangan positif sebesar 0,5593 dan pasangan negatif sebesar 0,5995. Nilai rata-rata pasangan negatif yang lebih tinggi menunjukkan bahwa model belum mampu membedakan pasangan teks yang memiliki hubungan semantik dengan pasangan yang tidak berkaitan secara optimal.

Sebaliknya, SimCSE-IndoBERT menghasilkan rata-rata *cosine similarity* sebesar 0,6767 pada pasangan positif dan 0,1355 pada pasangan negatif, dengan selisih sebesar 0,5412. Perbedaan tersebut menunjukkan bahwa proses *fine-tuning* berhasil menghasilkan representasi *embedding* yang lebih diskriminatif, sehingga pasangan teks yang relevan memperoleh nilai kemiripan yang lebih tinggi dibandingkan pasangan yang tidak relevan. Hasil ini menunjukkan bahwa SimCSE-IndoBERT lebih efektif dalam merepresentasikan kemiripan semantik dokumen tugas akhir dibandingkan model IndoBERT Base.

F. Pembentukan Representasi Teks

Pelatihan model menghasilkan model SimCSE-IndoBERT yang selanjutnya digunakan pada tahap pembentukan representasi teks. Proses *fine-tuning* memanfaatkan tiga jenis *positive pair*, yaitu *Self Pair*, *Weakly Supervised Pair*, dan *Augmented Pair*. Dari 1.445 data latih, proses pembentukan pasangan menghasilkan total 7.225 *positive pair* yang digunakan selama pelatihan. Ringkasan jumlah pasangan disajikan pada Tabel 10.

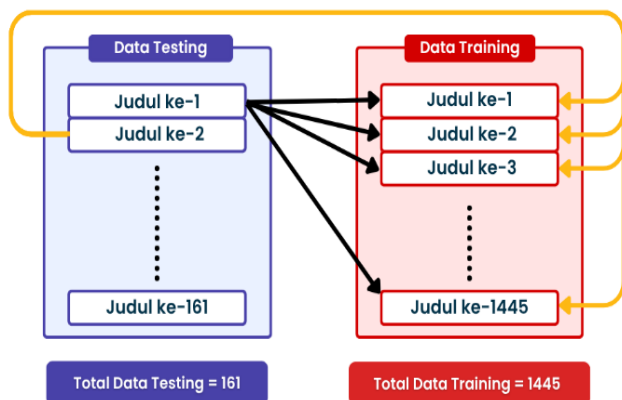
Tabel 10. Dimensi *Embedding* per Skenario

Skenario	Data Latih	Data Uji
Judul	1.445 × 768	161 × 768
Deskripsi	1.445 × 768	161 × 768
Gabungan Judul dan Deskripsi	1.445 × 768	161 × 768

Berdasarkan Tabel 10, seluruh skenario menghasilkan *embedding* berdimensi 768, sesuai dengan ukuran *hidden state* model IndoBERT. Jumlah baris pada setiap matriks mengikuti jumlah dokumen pada data latih dan data uji, yaitu masing-masing 1.445 dan 161 dokumen. Seluruh *embedding* kemudian disimpan sebagai representasi numerik dokumen yang digunakan pada tahap perhitungan *cosine similarity* untuk mengidentifikasi tingkat kemiripan semantik antar dokumen tugas akhir.

G. Perhitungan Kemiripan Cosine

Setelah seluruh dokumen direpresentasikan ke dalam bentuk *embedding*, dilakukan perhitungan *cosine similarity* untuk mengukur tingkat kemiripan semantik antara data uji dan data latih. Mekanisme perhitungan ditunjukkan pada Gambar 2, di mana setiap dokumen pada data uji dibandingkan dengan seluruh dokumen pada data latih menggunakan representasi vektor hasil *fine-tuning* IndoBERT.

Gambar 2. Mekanisme Perhitungan *Cosine Similarity*

Berdasarkan mekanisme tersebut, setiap dokumen pada data uji dibandingkan dengan 1.445 dokumen pada data latih. Dengan jumlah data uji sebanyak 161 dokumen, proses ini menghasilkan matriks kemiripan berukuran 161×1.445 , atau sebanyak 232.645 nilai *cosine similarity* pada setiap skenario pengujian.

Perhitungan dilakukan pada tiga skenario, yaitu menggunakan judul, deskripsi, serta gabungan judul dan deskripsi. Setiap proses menghasilkan matriks kemiripan yang berisi skor *cosine similarity* antara setiap data uji dan seluruh data latih.

H. Pengambilan Dokumen Termirip

Dalam pendekatan *dense retrieval*, representasi vektor digunakan untuk melakukan pencocokan antara *query* dan dokumen berdasarkan tingkat kemiripan dalam ruang representasi [20]. Setelah seluruh nilai *cosine similarity* diperoleh, sistem mengurutkan dokumen berdasarkan skor kemiripan dari yang tertinggi hingga terendah. Pada penelitian ini, lima dokumen dengan skor *cosine similarity* tertinggi dipilih sebagai hasil yang merepresentasikan dokumen dengan kedekatan semantik paling tinggi terhadap usulan tugas akhir yang diuji.

Sebagai contoh, digunakan satu data uji (*query*) dengan judul "Rancang Bangun Pemantau Berkurangnya Cairan Infus Berbasis Arduino dan IoT Klinik Azzaniyah Nurul Jadid" dan deskripsi "Alat ini berguna untuk memantau cairan infus yang ada di Klinik Azzaniyah Nurul Jadid sehingga memudahkan perawat atau dokter memantau kondisi cairan infus agar tidak sampai habis dan menghambat proses penyembuhan pasien." Berdasarkan *query* tersebut, sistem melakukan pencarian terhadap 1.445 dokumen pada data latih dan mengembalikan lima dokumen dengan skor *cosine similarity* tertinggi pada masing-masing skenario, yaitu judul, deskripsi, serta gabungan judul dan deskripsi. Hasil pengambilan dokumen termirip ditunjukkan pada Tabel 11, Tabel 12, dan Tabel 13. Perbedaan hasil pada setiap skenario menunjukkan pengaruh jenis masukan terhadap tingkat kemiripan yang dihasilkan oleh model dalam menemukan dokumen yang relevan.

Tabel 11. Hasil Pengambilan 5 Dokumen Termirip Pada Skenario Judul

Rank	Isi Teks	Skor
1	monitoring infus dan pulse heart rate berbasis iot	0,66
2	aplikasi sensor tekanan pada tabung oksigen berbasis iot di rumah sakit waluyo jati menggunakan hp android	0,59
4	rancang bangun sistem monitoring arus	0,57

5	tegangan listrik dan debit air pada pembangkit listrik tenaga mikrohidro berbasis internet of thing iot sistem monitoring smart klinik berbasis internet of things iot	0,56
---	---	------

Tabel 12. Hasil Pengambilan 5 Dokumen Termirip Pada Skenario Deskripsi

Rank	Isi Teks	Skor
1	Dengan menggunakan konsep jaringan internet of things kita dapat membuat sebuah sistem yang bisa memantau infus pada pasien sensor yang di ...	0,77
2	Aplikasi menggunakan sensor tekanan pada tabung oksigen berbasis iot ini memudahkan dokter dan perawat di ...	0,53
3	Alat ini dibuat berdasarkan latar belakang permasalahan yang ada pada lingkungan sekitar pondok pesantren nurul jadid salah satunya ialah ...	0,44
4	Klinik az-zainiyah adalah sebuah instansi kesehatan yang berada dibawah naungan pondok pesantren ...	0,43
5	Penelitian ini dilakukan berlandaskan dari penelitian sebelumnya yang menjelaskan bahwa untuk meningkatkan self manajemen ...	0,43

Tabel 13. Hasil Pengambilan 5 Dokumen Termirip Pada Skenario Gabungan Judul dan Deskripsi

Rank	Isi Teks	Skor
1	Monitoring infus dan pulse heart rate berbasis iot Dengan menggunakan konsep jaringan internet of things kita dapat kita dapat membuat sebuah sistem yang bisa ...	0,75
2	Aplikasi sensor tekanan pada tabung oksigen berbasis iot di ... Aplikasi menggunakan sensor tekanan pada tabung oksigen berbasis iot ini memudahkan dokter dan perawat di ...	0,54
3	Rancang bangun sistem pengendalian persediaan air ... Rancang bangun monitoring system otomasi water level tank dan water level tank dan water pump untuk ...	0,52
4	Rekam data self manajemen cairan pada pasien gagal ginjal kronis ... Penelitian ini dilakukan berlandaskan dari penelitian sebelumnya yang menjelaskan bahwa untuk meningkatkan self manajemen ...	0,49
5	Rancang bangun alat sensor pendeteksi kekeruhan air pada ... Alat ini dibuat berdasarkan latar belakang permasalahan yang ada pada lingkungan sekitar pondok pesantren nurul jadid salah satunya ialah ...	0,48

Berdasarkan hasil *Top-5 Similarity Retrieval*, ketiga skenario berhasil mengembalikan dokumen yang memiliki topik serupa dengan data uji. Pada seluruh skenario, dokumen yang berkaitan dengan monitoring infus berbasis Internet of Things (IoT) menempati peringkat pertama, menunjukkan bahwa model SimCSE-IndoBERT mampu mengidentifikasi keterkaitan semantik antara dokumen uji dan dokumen pada data latih.

Perbedaan terlihat pada skor *cosine similarity* yang dihasilkan. Skenario deskripsi memperoleh skor tertinggi sebesar 0,77, diikuti skenario gabungan judul dan deskripsi sebesar 0,75, sedangkan skenario judul menghasilkan skor 0,66. Hasil tersebut menunjukkan bahwa deskripsi mengandung informasi yang lebih lengkap dibandingkan judul sehingga menghasilkan representasi *embedding* yang lebih baik dalam menangkap hubungan semantik antar dokumen.

I. Evaluasi Model

Evaluasi dilakukan untuk mengukur kemampuan model dalam menemukan dokumen yang relevan pada tiga skenario, yaitu berdasarkan judul, deskripsi, serta gabungan judul dan deskripsi. Pengujian dilakukan menggunakan metrik *Precision@5*, *Recall@5*, dan *F1-Score@5* pada beberapa variasi *threshold cosine similarity* mulai dari 0,2 hingga 1,0.

Tabel 14. Ringkasan Nilai Evaluasi Tertinggi

Skenario	<i>P@5</i>	<i>R@5</i>	<i>F@5</i>	<i>Threshold Optimal</i>
Judul	0,9	0,51	0,65	0,5
Deskripsi	0,66	0,76	0,71	0,6
Gabungan	0,71	0,71	0,71	0,6

Ringkasan hasil evaluasi pada *threshold* terbaik tersebut disajikan pada Tabel 14.

Berdasarkan Tabel 14, skenario deskripsi dan gabungan judul serta deskripsi menghasilkan nilai *F1-Score@5* tertinggi sebesar 0,71, sedangkan skenario judul memperoleh nilai 0,65. Hasil tersebut menunjukkan bahwa informasi yang terdapat pada deskripsi mampu merepresentasikan konteks penelitian dengan lebih baik dibandingkan penggunaan judul saja sehingga meningkatkan efektivitas proses *retrieval*.

Skenario deskripsi menghasilkan nilai *Recall@5* tertinggi sebesar 0,76, yang menunjukkan kemampuan model menemukan lebih banyak dokumen relevan. Sementara itu, skenario gabungan judul dan deskripsi memberikan *Precision@5* yang lebih tinggi dibandingkan skenario deskripsi, yaitu sebesar 0,71, sehingga menghasilkan keseimbangan yang baik antara ketepatan dan kelengkapan hasil pencarian.

Berdasarkan hasil tersebut, nilai *threshold* 0,6 dipilih

sebagai konfigurasi yang digunakan pada sistem karena merupakan nilai optimal pada dua dari tiga skenario pengujian serta menghasilkan nilai *F1-Score@5* tertinggi secara keseluruhan. Hasil ini menunjukkan bahwa model SimCSE-IndoBERT mampu memberikan performa yang baik dalam mengidentifikasi dokumen tugas akhir yang memiliki kemiripan semantik sehingga dapat dimanfaatkan sebagai pendukung evaluasi awal usulan judul dan deskripsi tugas akhir.

J. Implementasi Model

Model SimCSE-IndoBERT yang telah melalui proses pelatihan diimplementasikan ke dalam aplikasi berbasis web untuk menguji kemampuannya dalam memberikan rekomendasi topik tugas akhir berdasarkan kemiripan semantik. Sistem menyediakan tiga mode pencarian, yaitu berdasarkan judul, deskripsi, serta gabungan judul dan deskripsi. Pada tahap implementasi, *embedding* data latih telah disimpan sebagai *pre-computed embedding*, sehingga sistem hanya melakukan pembentukan *embedding* terhadap teks masukan pengguna sebelum menghitung nilai *cosine similarity* terhadap seluruh data latih.

Gambar 3 menunjukkan halaman utama aplikasi yang digunakan untuk melakukan pengecekan kemiripan topik tugas akhir. Pengguna terlebih dahulu memilih mode pencarian yang diinginkan, kemudian memasukkan teks usulan tugas akhir sesuai mode yang dipilih. Setelah tombol Cek Kemiripan ditekan, sistem melakukan *preprocessing*, membentuk *embedding* menggunakan model SimCSE-IndoBERT, menghitung nilai *cosine similarity*, dan mengambil lima dokumen dengan tingkat kemiripan tertinggi sebagai hasil rekomendasi.

Sistem Scoring Judul Tugas Akhir Berbasis Indobert

Cek Kemiripan
Masukkan judul atau deskripsi penelitian

MODE PENGECEKAN

JUDUL TUGAS AKHIR

Sistem Informasi Surat Perintah Perjalanan Dinas di Dinas Komunikasi dan Informatika Kabupaten Situbondo Berbasis WEB

Hasil
Skor kemiripan dan keunikan penelitian

mode: Judul

"sistem informasi surat perintah perjalanan dinas di dinas komunikasi dan informatika kabupaten situbondo berbasis web"

Skor Keunikan Berdasarkan Nilai Judul Tertinggi

0.18

TOP 5 JUDUL TERMIRIP

NO	JUDUL	SKOR	LABEL
1	aplikasi surat perintah perjalanan dinas berbasis web studi kasus diskominfo kabupaten probolinggo	0.93	Sangat Mirip
2	aplikasi pengolahan surat perintah perjalanan dinas dengan global positioning system gps pada diskominfo kabupaten probolinggo berbasis web	0.86	Sangat Mirip
3	perancangan sistem monitoring surat perintah perjalanan dinas dengan mobile app android di universitas nurul jadid	0.8	Sangat Mirip
4	sistem informasi monitoring perjalanan dinas di universitas nurul jadid berbasis web	0.74	Mirip
5	rancang bangun aplikasi surat perintah perjalanan dinas spps menggunakan framework codeigniter di di universitas nurul jadid	0.74	Mirip

Gambar 3. Halaman Input dan Hasil Pengecekan Kemiripan

Sistem menyajikan skor keunikan beserta kategorinya, kemudian menampilkan lima dokumen dengan nilai *cosine similarity* tertinggi. Setiap dokumen disertai nilai kemiripan dan label klasifikasi sehingga pengguna dapat dengan mudah mengetahui tingkat kedekatan antara usulan tugas akhir dengan dokumen yang telah tersedia pada data latih. Untuk memudahkan interpretasi hasil pencarian, sistem mengelompokkan nilai *cosine similarity* ke dalam tiga kategori sebagaimana ditunjukkan pada Tabel 15.

Tabel 15. Klasifikasi Tingkat Kemiripan

Rentang Skor	Label
0,80 – 1	Sangat Mirip
0,60 – 0,79	Mirip
< 0,60	Tidak Mirip

Berdasarkan klasifikasi tersebut, dokumen dengan nilai *cosine similarity* yang semakin tinggi menunjukkan bahwa representasi semantik dokumen tersebut semakin dekat dengan usulan tugas akhir yang diuji. Sebaliknya, nilai kemiripan yang rendah mengindikasikan bahwa kedua dokumen memiliki keterkaitan semantik yang lebih kecil.

Selain menampilkan tingkat kemiripan, sistem juga menghitung skor keunikan berdasarkan nilai *cosine similarity* tertinggi yang diperoleh. Semakin tinggi nilai kemiripan terhadap dokumen pada data latih, semakin rendah skor keunikan yang dihasilkan. Rentang skor keunikan beserta kategorinya ditunjukkan pada Tabel 16.

Tabel 16. Klasifikasi Tingkat Keunikan

Rentang Skor	Label
0,75 – 1,00	Sangat Unik
0,50 – 0,74	Cukup Unik
0,25 – 0,49	Kurang Unik
0 – 0,24	Tidak Unik

Sebagai contoh, pada pengujian menggunakan judul "Sistem Informasi Surat Perintah Perjalanan Dinas di Dinas Komunikasi dan Informatika Kabupaten Situbondo Berbasis WEB", dokumen dengan nilai *cosine similarity* tertinggi memperoleh skor sebesar 0,93. Skor keunikan dihitung menggunakan normalisasi *Min-Max Scaling* yang ditunjukkan pada Persamaan (1), kemudian ditransformasikan menjadi skor keunikan menggunakan Persamaan (2).

$$S_{norm} = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (1)$$

$$Skor\ Keunikan = 1 - S_{norm} \quad (2)$$

Keterangan :

- S_{norm} = nilai kemiripan setelah normalisasi
- S = nilai *cosine similarity* tertinggi
- S_{min} = nilai *threshold* minimum yang ditetapkan
- S_{max} = nilai maksimum *cosine similarity*

Dengan mensubstitusikan nilai $S = 0,93$, $S_{min} = 0,60$ dan $S_{max} = 1,00$, diperoleh hasil perhitungan sebagai berikut.

$$S_{norm} = \frac{S - S_{min}}{S_{max} - S_{min}} = \frac{0,93 - 0,60}{1,00 - 0,60} = \frac{0,33}{0,4} = 0,825$$

Selanjutnya, berdasarkan Persamaan (2), skor keunikan dihitung sebagai berikut:

$$Skor\ Keunikan = 1 - S_{norm} = 1 - 0,825 = 0,175$$

Nilai tersebut dibulatkan menjadi 0,18, sehingga termasuk dalam kategori Tidak Unik sesuai dengan rentang pada Tabel 16. Hasil ini menunjukkan bahwa usulan tugas akhir memiliki tingkat kemiripan yang sangat tinggi terhadap salah satu dokumen pada data latih. Sebaliknya, apabila nilai *cosine similarity* tertinggi semakin rendah, maka skor keunikan yang dihasilkan akan semakin tinggi sehingga usulan tugas akhir dinilai lebih unik.

L. Pengujian Prototype

Prototype diuji menggunakan *Black Box Testing* dan *User Testing*. *Black Box Testing* bertujuan untuk memverifikasi bahwa seluruh fungsi sistem berjalan sesuai dengan kebutuhan fungsional yang telah dirancang, sedangkan *User Testing* bertujuan untuk mengevaluasi aspek kemudahan penggunaan (*usability*) sistem. Adapun hasil *Black Box Testing* ditunjukkan pada Tabel 17.

Tabel 17. Pengujian Black Box

No	Fungsi yang Diuji	Hasil
1	Interaktivitas form input sesuai mode pencarian	√
2	Validasi input kosong	√
3	Pemrosesan input yang mengandung karakter khusus	√
4	Penampilan hasil Top-5 dokumen termirip	√
5	Perhitungan dan penampilan skor serta label keunikan	√
6	Penampilan notifikasi proses sistem	√
7	Penyimpanan riwayat pengujian secara otomatis	√
8	Penampilan tabel dan detail riwayat pengujian	√
9	Ekspor riwayat pengujian ke format Excel	√

Hasil pengujian *Black Box* menunjukkan bahwa seluruh fungsi utama sistem berjalan sesuai dengan hasil yang diharapkan tanpa ditemukan kegagalan fungsional. Seluruh skenario pengujian, mulai dari proses input, pencarian dokumen termirip, perhitungan skor keunikan, hingga pengelolaan riwayat pengujian, berhasil dijalankan dengan baik.

Dengan terpenuhinya seluruh kebutuhan fungsional tersebut, pengujian dilanjutkan pada aspek penerimaan pengguna untuk mengukur sejauh mana sistem dapat diterima dan digunakan dengan baik oleh pengguna akhir. Adapun skala penilaian yang digunakan dalam kuesioner pengujian pengguna disajikan pada Tabel 18 berikut.

Tabel 18. Bobot Penilaian Skala Likert

Skala Jawaban	Bobot Nilai
Sangat Tidak Setuju (STS)	1
Tidak Setuju (TS)	2
Cukup (C)	3
Setuju (S)	4
Sangat Setuju (SS)	5

Selanjutnya, untuk menginterpretasikan hasil persentase kelayakan yang diperoleh dari kuesioner tersebut, digunakan kriteria kelayakan sistem sebagaimana ditunjukkan pada Tabel 19.

Tabel 19. Kriteria Persentase Kelayakan

Kriteria	Persentase
Sangat Tidak Layak (STL)	0% - 20%
Tidak Layak (TL)	21% - 40%
Cukup Layak (CL)	41% - 60%
Layak (L)	61% - 80%
Sangat Layak (SL)	81% - 100%

Berdasarkan skala penilaian dan kriteria tersebut, hasil pengisian kuesioner oleh responden kemudian direkapitulasi untuk memperoleh total skor setiap butir pertanyaan sebagaimana disajikan pada Tabel 20.

Tabel 20. Pengujian Penggunaan

No	Pernyataan	Total Skor
1	Aplikasi ini sangat mudah digunakan meskipun tanpa pelatihan khusus	42
2	Tabel hasil pencarian kemiripan teks di halaman utama sangat mudah dibaca	44
3	Informasi dan tata letak pada antarmuka website ditampilkan secara jelas dan terstruktur	43
4	Fitur formulir dinamis (<i>dynamic form</i>) sangat mempermudah dalam proses penginputan usulan tugas akhir	40
5	Proses pengecekan kemiripan teks berlangsung dengan sangat cepat.	42
6	Lima data tugas akhir teratas (Top-5) yang ditampilkan sangat relevan dengan teks yang diinputkan.	42
7	Label skor keunikan sangat membantu dalam proses evaluasi kelayakan usulan tugas akhir	42
8	Pesan notifikasi dari sistem (berhasil/gagal) ditampilkan dengan sangat jelas dan informatif	40
9	Fitur riwayat dan unduh laporan <i>Excel</i> sangat praktis	44
10	Secara keseluruhan, sistem <i>scoring</i> tugas akhir ini sangat membantu dalam proses evaluasi kelayakan usulan tugas akhir	43
Total Skor		422

Berdasarkan Tabel 20, diperoleh total skor sebesar 422 dari skor ideal 500 yang dihitung berdasarkan 10 responden, 10 pertanyaan, dan skor maksimum 5 untuk setiap pertanyaan. Persentase kelayakan sistem dihitung menggunakan Persamaan (3).

$$\text{Persentase Kelayakan} = \frac{\text{Total Skor}}{\text{Skor Ideal}} \times 100\% \quad (3)$$

Dengan mensubstitusikan nilai total skor dan skor ideal ke dalam Persamaan (3), diperoleh:

$$\text{Persentase Kelayakan} = \frac{422}{500} \times 100\% = 84,40\%$$

Berdasarkan kriteria kelayakan pada Tabel 19, nilai 84,40% termasuk dalam kategori Sangat Layak. Hasil tersebut menunjukkan bahwa pengguna memberikan penilaian positif terhadap aplikasi. Dengan demikian, *prototype* aplikasi memiliki tingkat *usability* yang baik dan layak digunakan sebagai alat bantu evaluasi awal usulan tugas akhir mahasiswa.

IV. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengoptimalkan model IndoBERT menggunakan pendekatan SimCSE untuk mengukur kemiripan semantik usulan tugas akhir. Hasil fine-tuning meningkatkan kemampuan model dalam membedakan pasangan teks yang relevan dan tidak relevan, yang ditunjukkan oleh peningkatan selisih rata-rata nilai cosine similarity antara pasangan positif dan negatif dari 0,040 pada IndoBERT Base menjadi 0,5412 pada SimCSE-IndoBERT.

Berdasarkan evaluasi menggunakan metrik Precision@5, Recall@5, dan F1-Score@5, skenario deskripsi dan gabungan memberikan performa terbaik dengan F1-Score@5 sebesar 0,71 pada threshold 0,6. Model kemudian berhasil diimplementasikan ke dalam aplikasi berbasis web untuk mendukung evaluasi awal usulan tugas akhir, dengan hasil pengujian menunjukkan seluruh fungsi sistem berjalan dengan baik serta memperoleh tingkat penerimaan pengguna sebesar 84,40% yang termasuk kategori Sangat Layak.

Hasil penelitian menunjukkan bahwa pendekatan SimCSE-IndoBERT efektif dalam mendukung identifikasi kemiripan semantik dan penilaian keunikan usulan tugas akhir. Penelitian selanjutnya dapat meningkatkan kualitas sistem dengan memperluas cakupan dataset, melibatkan pakar domain dalam penentuan ground truth pada pasangan positif dan negatif, mengeksplorasi pendekatan Supervised SimCSE atau model representasi kalimat yang lebih mutakhir, serta memanfaatkan informasi dokumen yang lebih lengkap agar penilaian kemiripan dan keunikan menjadi lebih akurat.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan dalam pelaksanaan penelitian ini. Ucapan terima kasih secara khusus disampaikan kepada dosen pembimbing Program Studi S1 Teknik Informatika Universitas Nurul Jadid atas bimbingan, arahan, kritik, dan saran yang sangat berarti selama proses penelitian hingga penyusunan artikel ini. responden yang telah berpartisipasi dalam pengujian sistem, serta keluarga dan rekan-rekan yang senantiasa memberikan doa, dukungan, dan motivasi sehingga penelitian ini dapat diselesaikan dengan baik.

DAFTAR PUSTAKA

- [1] E. Abdul Goffar And D. D. Wazaumi, "Pemodelan Sistem Rekomendasi Judul Tugas Akhir Mahasiswa Berdasarkan Personalisasi Mahasiswa Tingkat Akhir Dan Implementasi Model Berbasis Situs Web," *Infotech: Jurnal Informatika & Teknologi*, Vol. 7, No. 1, Pp. 106–114, Jun. 2026, Doi: 10.37373/Infotech.V7i1.2139.
- [2] N. Fadhlila, C. Apriyanti, And I. Puspitasari, "The Difficulties Of Students In Writing A Thesis At Skip Pgri Pacitan In Academic Year 2022/2023," 2023. [Online]. Available: <https://Repository.Stkippacitan.Ac.Id/>
- [3] M. W. Malo And R. S. Kaka, "Analisis Kesulitan Penyelesaian Skripsi Pada Mahasiswa Program Studi Teknik Informatika Universitas Stella Maris Sumba," *Jurnal Pendidikan Dan Pemikiran*, Vol. 20, No. 1, 2025, Doi: 10.55558/Alihda.V20i1.225.

- [4] L. Bornmann, R. Haunschild, And R. Mutz, "Growth Rates Of Modern Science: A Latent Piecewise Growth Curve Approach To Model Publication Numbers From Established And New Literature Databases," *Humanit. Soc. Sci. Commun.*, Vol. 8, No. 1, Dec. 2021, Doi: 10.1057/S41599-021-00903-W.
- [5] E. Elhosary And O. Moselhi, "Evaluating Natural Language Processing Algorithms For Improved Hazard And Operability Analysis," *Geodata And Ai*, P. 100026, 2025, Doi: 10.1016/J.Geoai.2025.100026.
- [6] I. Hakiki, F. Atqiya, And A. Suryan, "Implementasi Model Sentence-Bert Untuk Deteksi Plagiarisme Pada Karya Tulis Ilmiah Psikologi Berbahasa Indonesia.," *Cerdika: Jurnal Ilmiah Indonesia*, Vol. 5, No. 12, 2025, Doi: 10.59141/Cerdika.V5i12.2889.
- [7] A. P. Putra, D. P. S. Putri, And A. A. K. A. C. Wiranatha, "Scientific Paper Recommendation System: Application Of Sentence Transformers And Cosine Similarity Using Arxiv Data," *Journal Of Applied Informatics And Computing*, Vol. 9, No. 4, Pp. 1374–1382, 2025, Doi: 10.30871/Jaic.V9i4.9766.
- [8] F. Koto, A. Rahimi, J. H. Lau, And T. Baldwin, "Indolem And Indobert: A Benchmark Dataset And Pre-Trained Language Model For Indonesian Nlp," *Arxiv Preprint Arxiv:2011.00677*, 2020.
- [9] H. Choi, J. Kim, S. Joe, And Y. Gwon, "Evaluation Of Bert And Albert Sentence Embedding Performance On Downstream Nlp Tasks," Jan. 2021, Doi: [Http://Arxiv.Org/Abs/2101.10642](http://arxiv.org/abs/2101.10642).
- [10] H. Murfi, T. Gowandi, G. Ardanawari, And S. Nur-Rohmah, "Bert-Based Combination Of Convolutional And Recurrent Neural Network For Indonesian Sentiment Analysis," 2022. Doi: <https://doi.org/10.48550/Arxiv.2211.05273>.
- [11] N. Reimers And I. Gurevych, "Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks," 2019. Doi: 10.18653/V1/D19-1410.
- [12] K. Ethayarajh, "How Contextual Are Contextualized Word Representations? Comparing The Geometry Of Bert, Elmo, And Gpt-2 Embeddings," Nov. 2019. Accessed: Apr. 05, 2026. [Online]. Available: <https://aclanthology.org/D19-1006/>
- [13] T. Gao, X. Yao, And D. Chen, "Simcse: Simple Contrastive Learning Of Sentence Embeddings," 2021. Accessed: Jun. 24, 2026. [Online]. Available: <https://aclanthology.org/2021.emnlp-main.552/>
- [14] M. Henderson *Et Al.*, "Efficient Natural Language Response Suggestion For Smart Reply," May 2017, Doi: [Doi.Org/10.48550/Arxiv.1705.00652](https://doi.org/10.48550/Arxiv.1705.00652).
- [15] V. Karpukhin *Et Al.*, "Dense Passage Retrieval For Open-Domain Question Answering," 2020. Doi: 10.48550/Arxiv.2004.04906.
- [16] X. Yuan, Z. Yang, Y. Wang, J. Zhao, And K. Liu, "Improving Zero-Shot Llm Re-Ranker With Risk Minimization," 2024. Doi: 10.18653/V1/2024.Emnlp-Main.997.
- [17] Z. Wang, J. Dou, And Y. Zhang, "Unsupervised Sentence Textual Similarity With Compositional Phrase Semantics," 2022. Doi: <https://aclanthology.org/2022.coling-1.441/>.
- [18] A. Levy, B. R. Shalom, And M. Chalamish, "A Guide To Similarity Measures And Their Data Science Applications," *J. Big Data*, Vol. 12, No. 1, Dec. 2025, Doi: 10.1186/S40537-025-01227-1.
- [19] A. A. Sari, N. Arifin, And C. N. Insani, "Klasifikasi Penyakit Buah Kakao Berbasis Citra Menggunakan Transfer Learning Dengan Model Mobilenetv2," Vol. 11, No. 1, 2026, Doi: 10.51876/Simtek.V11i1.1761.
- [20] W. X. Zhao, J. Liu, R. Ren, And J. R. Wen, "Dense Text Retrieval Based On Pretrained Language Models: A Survey," *Acm Trans. Inf. Syst.*, Vol. 42, No. 4, Feb. 2024, Doi: 10.1145/3637870.