Jurnal PETISI, Vol. 07, No. 01, Januari 2026
e-ISSN: 2721-6276

Algoritma *Deep Learning* untuk Mengukur Tingkat Kesesuaian Rumusan dan *Asesmen* Capaian Pembelajaran Mata Kuliah

Wuryanto[1], Husni Thamrin[2], Jan Wantoro[3]
Magister Informatika, Fakultas Komunikasi dan Informatika
Universitas Muhammadiyah Surakarta
husni.thamrin@ums.ac.id

Abstrak: Penelitian ini mengamati kinerja algoritma deep learning untuk mengukur keselarasan antara pertanyaan ujian dengan capaian pembelajaran matakuliah. Keselarasan kedua teks menggambarkan kewajaran proses pembelajaran, sehingga proses terbukti telah berjalan sesuai rencana pembelajaran. Secara umum, proses pembelajaran dimulai dari penyusunan rencana pembelajaran, yang menghasilkan dokumen yang mengandung capaian pembelajaran (learning outcome). Selanjutnya proses belajar mengajar dijalankan baik secara daring ataupun luring, dan capaian pembelajaran diukur dengan metode asesmen yang mengukur secara substantif tingkat pencapaian pembelajaran. Riset dimulai dengan pengumpulan data dan *preprocessing*, dan berlanjut dengan menghitung tingkat kesesuaian menggunakan tiga model deep learning. Ketiga model yang diuji adalah sentence-transformer (model SBERT), denaya/indoSBERT-large (model Denaya), and cahya/bert-base-indonesian-1.5G (model Cahya). Penelitian ini menunjukkan bahwa model Cahya memiliki kinerja terbaik dengan nilai akurasi 0.92 dan *f1-score* 0.88 dibanding dengan evaluasi oleh manusia. Studi ini menunjukkan bahwa algoritma deep learning dapat diterapkan untuk mereview teks soal dalam kegiatan ujian agar selaras dengan capaian pembelajaran.
Kata Kunci : deep learning, capaian pembelajaran, asesmen, kinerja model, natural language processing.

Abstract: *The study examines the performance of deep learning algorithms in measuring the alignment level of exam questions and course learning outcomes. The alignment of the two documents indicates the appropriateness of the learning process, whereby the process proves to have run following the plan. Generally, a learning process starts with course planning, which produces a document containing the learning outcomes. Then, teaching-learning is carried out in educational settings either online or offline, and learning outcomes are measured using assessment methods that must substantively meter the level of learning achievement. The research starts with data collection and preprocessing and proceeds with calculating the level of document alignment using three deep-learning algorithms. Performance evaluation uses expert judgment as the benchmark. The research investigates three deep learning models. They include sentence-transformer (SBERT model), denaya/indoSBERT-large (Denaya model), and cahya/bert-base-indonesian-1.5G (Cahya model). Our study shows that the Cahya model performs best showing an accuracy of 0.92 and a weighted f1-Score at 0.88 against human reviewers' evaluation. The study suggests that the Deep Learning algorithm can be implemented to review question texts in examination settings to align with the course learning outcomes.*
Keywords: deep learning, course learning outcome, assessment, model performance, natural language processing

1. Pendahuluan

Secara umum, kualitas pendidikan sangat dipengaruhi oleh kualitas proses pembelajaran, yang ditentukan oleh keterpaduan berbagai komponen pendidikan. Kurikulum menjadi komponen strategis karena berfungsi sebagai seperangkat rencana dan pengaturan mengenai tujuan, isi, bahan ajar, serta metode pembelajaran yang digunakan sebagai pedoman dalam penyelenggaraan pembelajaran untuk mencapai tujuan pendidikan tertentu(Government of the Republic of Indonesia. (2003), 2003). Seiring meningkatnya tuntutan mutu pendidikan, berkembang berbagai model kurikulum, salah satunya adalah kurikulum berbasis luaran atau *outcome-based education* (OBE), yang menekankan pencapaian kemampuan peserta didik setelah mengikuti proses pembelajaran. Pendekatan OBE terbukti mampu meningkatkan kualitas pembelajaran dan menghasilkan lulusan yang lebih kompeten (Nurjannah et al., 2021; Subadi et al., 2022).

Implementasi OBE dimulai dari perumusan Capaian Pembelajaran Lulusan (CPL) yang selanjutnya diturunkan ke dalam Capaian Pembelajaran Mata Kuliah (CPMK). CPMK kemudian dijabarkan dalam Rencana Pembelajaran Semester (RPS), yang memuat materi perkuliahan serta rencana asesmen untuk mengukur ketercapaian(Davis, 2003). RPS yang disusun dengan baik memungkinkan dosen merancang pembelajaran secara sistematis dan terarah, sedangkan RPS yang kurang berkualitas berpotensi menurunkan efektivitas pembelajaran (Devasis Pradhan, 2021). Oleh karena itu, ketersediaan dan penerapan RPS yang berkualitas menjadi faktor penting dalam meningkatkan kelulusan tepat waktu dan mutu lulusan perguruan tinggi(Abas & Imam, 2016; Rojak et al., 2022).

Namun demikian, belum semua dosen memahami konsep OBE secara utuh, sehingga RPS yang disusun sering kali belum selaras antara CPL, CPMK, aktivitas pembelajaran, dan instrumen asesmen. Untuk mengatasi hal tersebut, berbagai perguruan tinggi menerapkan proses *peer review* terhadap RPS dan instrumen asesmen, baik melalui telaah tertulis berbasis sistem informasi maupun presentasi di tingkat program studi. Proses review manual yang dilakukan secara rutin memerlukan sumber daya yang besar dan berpotensi menghasilkan penilaian yang tidak konsisten antar reviewer.

Permasalahan tersebut mendorong perlunya alternatif solusi yang lebih efisien dan konsisten. Perkembangan kecerdasan buatan, khususnya deep learning dalam pemrosesan bahasa alami (NLP), membuka peluang untuk membantu tugas-tugas analitis yang bersifat repetitif (Kuppili et al., 2020a; Mukherjee et al., 2022). Berbagai kajian NLP telah mengembangkan metode untuk menilai kesamaan tema dan makna antar dokumen (Reimers & Gurevych, 2019a; Thamrin & Sabardila, 2016). yang memungkinkan evaluasi konsistensi isi dokumen pembelajaran seperti RPS dan instrumen asesmen.

Berbagai pendekatan NLP berbasis *machine learning* telah diterapkan pada tugas peringkasan dokumen(Goularte et al., 2019a), pengelompokan dokumen(Dengel & Dubiel, 1995), klasifikasi teks (Thamrin et al., 2021; Zade & Ajani, 2022), hingga penerjemahan(T S et al., 2019). Pada beberapa kasus, kinerja sistem berbasis kecerdasan buatan bahkan mampu menyamai atau melampaui kemampuan manusia dalam tugas klasifikasi tertentu.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengukur tingkat kesesuaian antara rumusan CPMK dan instrumen asesmen pembelajaran dengan memanfaatkan pendekatan *deep learning* berbasis NLP. Secara khusus, penelitian ini mengevaluasi kinerja tiga model berbasis BERT, yaitu sentence-transformer/bert-base-nli-mean-tokens (SBERT), denaya/indoSBERT-large, dan cahya/bert-base-indonesian-

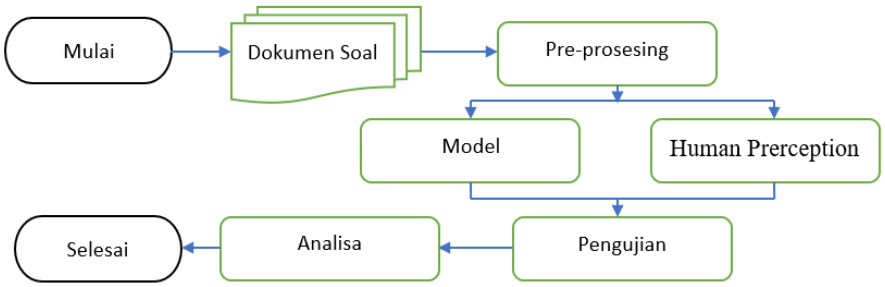
1.5G, dalam mengklasifikasikan kesesuaian teks CPMK dan soal ujian berdasarkan kemiripan semantik.

Kontribusi ilmiah penelitian ini terletak pada penerapan dan evaluasi komparatif model *deep learning* berbasis BERT berbahasa Indonesia untuk menilai kesesuaian CPMK dan instrumen asesmen, yang masih relatif terbatas dikaji dalam konteks pendidikan tinggi di Indonesia. Penelitian ini menggunakan dokumen autentik pendidikan tinggi, melibatkan penilaian pakar sebagai pembanding, serta mengkaji pengaruh *pretraining* model bahasa Indonesia berskala besar terhadap kinerja klasifikasi kesesuaian pembelajaran. Selain itu, penelitian ini memberikan kontribusi praktis dengan membuka peluang pemanfaatan kecerdasan buatan sebagai alat bantu awal (*screening tool*) dalam sistem penjaminan mutu akademik, khususnya untuk mendukung proses review RPS dan asesmen secara lebih efisien dan konsisten.

2. Metode Penelitian

2.1 Tahapan Penelitian

Penelitian ini merupakan penelitian eksperimental dengan pendekatan komputasional yaitu yang menggabungkan elemen eksperimental dengan penggunaan teknik komputasional, khususnya dalam konteks ilmu komputer dan pengolahan data(Maharani et al., 2020). Tahapan yang dilakukan dalam penelitian dapat dilihat pada gambar 1



Gambar 1. Tahapan Penelitian

2.1.1 Pengumpulan Data

Data yang digunakan dalam penelitian adalah dokumen RPS dan teks soal yang diujikan pada mahasiswa di Universitas Muhammadiyah Surakarta pada semester genap 2021/2022 sampai dengan genap 2022/2023. Dokumen soal dapat dikoleksi karena terdapat proses review sebelum pelaksanaan ujian tengah semester dan ujian akhir semester. Tabel 1 adalah data mentah dokumen soal yang diperoleh di 10 fakultas dalam bentuk PDF dan DOCX.

Tabel 1 Daftar Dokumen Soal

Fakultas	Jumlah Dokumen
Fakultas Keguruan dan Ilmu Pendidikan	2380
Fakultas Ekonomi dan Bisnis	738
Fakultas Hukum	292
Fakultas Teknik	1760
Fakultas Geografi	233
Fakultas Psikologi	100
Fakultas Agama Islam	599
Fakultas Ilmu Kesehatan	1026

Fakultas	Jumlah Dokumen
Fakultas Farmasi	208
Fakultas Komunikasi dan Informatika	927
Total	**8263**

Dari total 8.263 dokumen soal yang terkumpul, terdiri atas 7.117 dokumen berformat PDF dan 1.142 dokumen berformat DOCX. Dalam penelitian ini, hanya dokumen berformat PDF yang diproses lebih lanjut. Keputusan ini didasarkan pada pertimbangan metodologis terkait konsistensi struktur dokumen dan keandalan proses ekstraksi teks.

Dokumen PDF yang digunakan telah mengikuti template soal institusional yang seragam, sehingga memungkinkan identifikasi otomatis bagian rumusan CPMK dan butir soal secara konsisten. Konsistensi struktur ini menjadi prasyarat penting untuk memastikan bahwa teks CPMK dan teks soal yang diekstraksi benar-benar merepresentasikan pasangan semantik yang valid untuk analisis kesesuaian.

Sebaliknya, dokumen berformat DOCX menunjukkan variasi format yang tinggi, baik dalam tata letak, penamaan bagian, maupun penempatan rumusan CPMK dan soal. Variasi ini berpotensi menimbulkan kesalahan ekstraksi dan ketidaktepatan pemetaan CPMK–soal, yang dapat berdampak negatif terhadap validitas pengukuran kesamaan semantik. Oleh karena itu, dokumen DOCX tidak disertakan dalam tahap pemrosesan data pada penelitian ini.

Meskipun demikian, pengolahan dokumen DOCX dengan pendekatan normalisasi format atau deteksi struktur berbasis pembelajaran mesin merupakan peluang pengembangan yang relevan dan akan dipertimbangkan dalam penelitian lanjutan.

UJIAN TENGAH SEMESTER GASAL 2023/2024
ODD MIDTERM EXAM 2023/2024

FAKULTAS (Faculty) : KIP (Teacher Training And Education)			
JURUSAN (Department) : PEND. AKUNTANSI (Accounting Education)			
Mata Uji - Course	Pancasila	Hari/Tanggal - Day/Date	Selasa / 31 Oktober 2023
Smt/Kelas - Class	1 / A	Jam ke - Session	1
Penguji - Examiner	1. Dra. Sundari, M.Hum	Waktu - Duration	90 Menit
Petunjuk - Guidance:			
1.			
2.			
Capaian Pembelajaran Mata Kuliah - Course Learning Outcomes (CPMK - CLO):			
1.			
2.			

Soal Tipe A - Type A Questions (contoh jika ada beberapa tipe soal)

No	Soal - Questions	Nilai - Score	CPMK - CLO
1.			
2.			
3.			
dst.			

Gambar 2. Template Soal

2.1.2 Pra Proses Data

Preprocessing adalah langkah yang penting dalam penelitian ini. Praproses merupakan tahapan untuk menghilangkan simbol, karakter, dan tanda baca yang tidak relevan, serta memperbaiki kesalahan tulis yang dapat memengaruhi hasil pengolahan data (Nguyen et al., 2019a). *Case folding*, *tokenization, filtering*, dan *stemming* adalah beberapa tahapan pra proses yang digunakan dalam pengolahan bahasa alami terhadap dokumen bertipe teks. Pemeriksaan ulang data juga dilakukan pada tahap *preprocessing*, termasuk menghilangkan redundansi, *outlier*, dan nilai null (data kosong). Pemeriksaan ulang dilakukan untuk memastikan data yang diproses adalah

data yang "bersih", sehingga dapat memastikan bahwa hasil perhitungan algoritma akan memberikan hasil yang sesuai.

2.1.3 *Human Perception*

Pada tahapan ini peneliti mencoba menggali data dari persepsi manusia tentang kesesuaian rumusan CPMK dan teks soal. *Human Perception* menjadi benchmark dalam penentuan kesesuaian antar dokumen pembelajaran. Proses penggalian data melibatkan 22 dosen dari berbagai bidang ilmu. Setiap pasang teks rumusan CPMK dan soal dinilai oleh dua dosen. Dosen memilih salah satu dari opsi yang sudah ditetapkan yaitu Tidak Sesuai, Kurang Sesuai, Sesuai, Sangat Sesuai, dan Skip. Opsi Skip dapat dipilih oleh dosen jika terdapat kesalahan pada teks soal atau rumusan CPMK sehingga maknanya tidak dapat dimengerti. Dari 904 pasangan teks soal dan rumusan CPMK, didapatkan 775 pasang data yang terbaca dan dapat dimengerti dan 129 pasang tidak saling berkaitan.

2.1.4 Model

Model berbasis *embedding* kata dan kalimat memiliki peran krusial dalam menganalisis teks. Word Embedding-Based Models seperti Word2Vec, GloVe, dan FastText mengkodekan kata-kata menjadi vektor, memungkinkan perbandingan makna kata dengan mengukur kesamaan vektor antar kata. Di sisi lain, Sentence Embedding-Based Models seperti Universal Sentence Encoder (USE), BERT(Mutinda et al., 2021), dan S-BERT menciptakan representasi vektor untuk kalimat, menangkap konteks dan kemiripan makna di antara kalimat.

Penelitian ini mengamati model BERT dan S-BERT. S-BERT atau Siamese BERT dikembangkan dari dua buah model BERT yang dilatih secara bersamaan, tekniknya serupa dengan BERT, namun dilatih khusus untuk mengekstrak informasi dari kalimat(Devi & Suadaa, 2022). Ketika digunakan untuk mengevaluasi kesesuaian antara CPMK dan asesmen, Siamese BERT memperoleh representasi vektor untuk keduanya, memungkinkan pengukuran kesamaan makna.

Pilihan algoritma tergantung pada kebutuhan spesifik, ukuran dataset, dan kompleksitas makna yang dibutuhkan. Model seperti SBERT (*sentence-bert/bert-base-nli-mean-tokens*), yang menggunakan Siamese BERT untuk menghasilkan representasi vektor dari kalimat dengan mengambil rata-rata token, bertujuan untuk memperoleh representasi makna kalimat secara keseluruhan(Kuppili et al., 2020a)(Devi & Suadaa, 2022). Dengan representasi ini, perbandingan makna antar kalimat dapat dilakukan dengan metrik kesamaan vektor, memungkinkan analisis efisien dalam ruang vektor yang telah dilatih dengan data besar.

Selain menggunakan model SBERT, penelitian ini menggunakan model Denaya (*denaya/indoSBERT-large*) dan model Cahya (*cahya/bert-base-indonesian-1.5G*). Model Denaya adalah model dengan prinsip kerja yang memetakan kalimat dan paragraf ke ruang vektor pada 256 dimensi dan disempurnakan dengan menggunakan skema jaringan siam yang terinspirasi oleh SBERT(Diana, 2023). Model ini disempurnakan dengan dataset STS (2012-2016) yang telah diterjemahkan ke dalam bahasa Indonesia. Sedangkan model Cahya (*cahya/bert-base-indonesian-1.5G*) adalah model berbasis BERT yang telah dilatih sebelumnya dengan Wikipedia bahasa Indonesia dan surat kabar berbahasa Indonesia dengan menggunakan tujuan pemodelan bahasa terselubung (*masked language modeling*/MLM). Model ini telah dilatih dengan 522MB Wikipedia bahasa Indonesia dan 1GB surat kabar berbahasa Indonesia. Teks-teks tersebut di-cropping dan diberi token menggunakan WordPiece dan ukuran kosakata sebesar 32.000(Luthfi et al., 2021a).

2.1.5 Evaluasi

Evaluasi terhadap kinerja model atau algoritma merupakan parameter kritis dalam menentukan efektivitas algoritma. Dalam konteks penelitian ini, hasil perhitungan algoritma dibandingkan dengan persepsi pengguna untuk melihat apakah sesuai atau tidak sesuai. Pendekatan ini tergolong klasifikasi dan kinerjanya biasa digambarkan dengan confusion matrix. Confusion matrix untuk klasifikasi biner memiliki empat elemen utama yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) (Visa et al., 2011). Berdasarkan komponen tersebut dapat dihitung accuracy, precision, dan recall.

a. Accuracy menunjukkan seberapa sering model memberikan prediksi yang benar secara keseluruhan terhadap seluruh kasus yang diuji. Akurasi dapat diekspresikan dengan persamaan (1).

$$Akurasi = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

b. Precision menunjukkan seberapa tepat model dalam membuat prediksi positif. Presisi mengukur seberapa banyak dari prediksi positif yang sebenarnya benar dengan menghindari nilai positif palsu. Presisi dapat diekspresikan dengan persamaan (2).

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

c. Recall atau Sensitivitas menunjukkan seberapa baik model dapat mendeteksi nilai yang sebenarnya positif. Recall dihitung mengikuti persamaan (3).

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

Presisi dan recall bersifat tradeoff di mana kenaikan nilai satu metrik biasanya berarti penurunan nilai metrik yang lain. F-Score adalah metrik gabungan yang mempertimbangkan keseimbangan antara presisi dan recall. Jika presisi dianggap lebih penting, bobot presisi dapat dinaikkan dan sebaliknya. F-Score menjadi metrik yang lebih tepat dibanding akurasi jika terdapat ketidakseimbangan kelas, di mana data untuk satu kelas lebih dominan daripada yang lain. Untuk bobot antara presisi dan recall dianggap setara, maka F-Score dapat dihitung dengan persamaan (4).

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

3. Hasil Penelitian

3.1 Pre-Processing Data

Dokumen soal bertipe PDF sebanyak 7117 dilakukan proses *preprocessing* dengan ekstraksi dokumen. Ekstraksi dokumen menggunakan *tools* dengan *library python pdfplumber. Tools* tersebut didesain untuk membaca CPMK dan pertanyaan (soal) serta hubungan keduanya. Jika format dokumen tidak sesuai dengan format yang disepakati, maka CPMK dan soal tidak bisa terbaca dengan benar. Dari total jumlah dokumen bertipe *pdf* yang digunakan 7117, diperoleh data pertanyaan yang terhubung dengan CPMK sebanyak 904 baris data. Tabel 2 memperlihatkan distribusi data untuk masing-masing fakultas di tempat penelitian ini berjalan.

Tabel 2. Jumlah Data Pertanyaan

Fakultas	Pertanyaan
Fakultas Keguruan dan Ilmu Pendidikan	245
Fakultas Ekonomi dan Bisnis	93
Fakultas Hukum	22
Fakultas Teknik	338
Fakultas Geografi	10
Fakultas Agama Islam	54
Fakultas Ilmu Kesehatan	30
Fakultas Komunikasi dan Informatika	112
Total	**904**

3.2 Pemrosesan Data

Setelah diperoleh data teks CPMK dan teks pertanyaan dari masing-masing soal, tahap selanjutnya adalah pemrosesan data. Kesesuaian teks soal dengan teks CPMK diukur berdasarkan tingkat kemiripan dari kedua teks. Perhitungan similaritas dilakukan dengan model BERT (Bidirectional Encoder Representations from Transformer)(Agirre et al., 2016). Kesesuaian ditentukan berdasarkan nilai similaritas dengan threshold pada level 0.5. Artinya nilai similaritas < 0.5 dikategorikan tidak sesuai dan similaritas lebih dari atau sama dengan 0.5 dikategorikan sesuai.

Untuk keperluan evaluasi penentuan kesesuaian antara teks soal dan CPMK, dilakukan pengambilan data (survei) penilaian kesesuaian teks soal dan CPMK kepada pengguna manusia. Beberapa kendala terkait keterbacaan soal menyebabkan data yang dapat diolah berkurang menjadi 763.

3.2.1 Model BERT

Proses perhitungan similaritas menggunakan tiga model BERT yaitu *sentence-transformer*/*bert-base-nli-mean-tokens* (SBERT), *denaya*/*indoSBERT-large* (Denaya) dan *cahya/bert-base-indonesian-1.5G* (Cahya). Berdasarkan perhitungan similaritas antara teks CPMK dan teks soal menggunakan vektor *embeddings* dari model SBERT didapat skor similaritas di bawah 0.5 sebanyak 87 pasang teks dan 676 pasang teks memiliki skor similaritas lebih besar atau sama dengan 0.5. Penggunaan model Denaya menghasilkan skor similaritas di bawah 0.5 sebanyak 349 pasang teks dan 414 pasang teks dengan skor lebih besar atau sama dengan 0.5. Sedangkan penerapan model *cahya/bert-base-indonesian-1.5G* menghasilkan skor similaritas di bawah 0.5 hanya 5 pasang dan skor lebih besar atau sama dengan 0.5 sebanyak 758 pasang.

Hasil perhitungan similaritas dikategorikan menjadi dua kelas yaitu: Tidak Sesuai (TS) dan Sesuai (S) dengan angka 0.5 sebagai threshold, sebagaimana dapat dilihat pada tabel 3.

Tabel 3. Klasifikasi nilai

Label	Range Nilai
Tidak Sesuai (TS)	0 - 0.5
Sesuai (S)	0.5 - 1

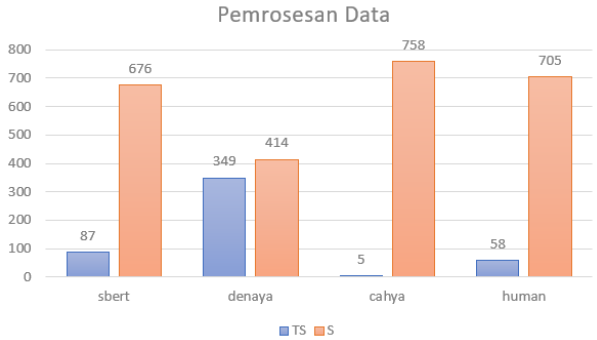
3.2.2 *Human Perception*

Persepsi manusia merupakan tolok ukur terhadap kinerja algoritma. Data persepsi manusia berupa penilaian terhadap kesesuaian antara teks soal dan CPMK. Pengambilan data persepsional dilakukan dengan melibatkan 22 dosen, terdiri dari 6 dosen Fakultas Teknik, 6 dosen Fakultas Keguruan dan ilmu pendidikan, 2 dosen Fakultas Kesehatan, 2 dosen Fakultas Agama Islam, 2 dosen Fakultas Ekonomi dan Bisnis, dan 4 dosen Fakultas Komunikasi dan Informatika.

Penilaian dilakukan dengan membuat aplikasi survei seperti pada gambar 2 di mana dosen memilih opsi Tidak Sesuai (TS), Kurang Sesuai (KS), Sesuai (S), dan Sangat Sesuai (SS). Opsi kelima yaitu Skip diberikan untuk mengantisipasi teks soal atau CPMK yang tidak terbaca atau teks soal yang dirasa tidak lengkap sehingga kesesuaiannya tidak dapat dinilai. Selanjutnya dalam analisis, dilakukan pemampatan jumlah kelas menjadi dua saja yaitu Sesuai dan Tidak Sesuai agar hasil penilaian persepsional dapat dibandingkan dengan hasil klasifikasi menggunakan algoritma.

Gambar 3. Formulir penilaian kesesuaian teks soal dan CPMK

Hasil klasifikasi menggunakan tiga model yaitu SBERT, Denaya, dan Cahya beserta human perception dapat dilihat pada gambar 3. Hasil klasifikasi berdasarkan data persepsional menunjukkan adanya ketidakseimbangan data untuk kedua kelas. Jumlah data yang masuk ke kelas Sesuai mencapai lebih dari 90% data yang ada.



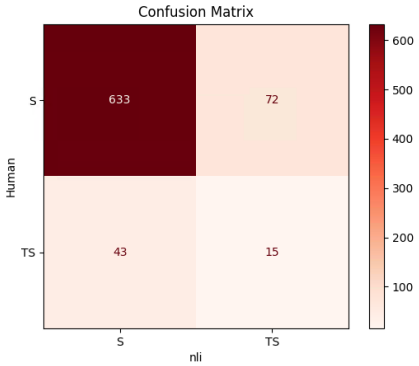
Gambar 4. Hasil klasifikasi ketiga model dan human perception

3.3 Evaluasi

Hasil klasifikasi menggunakan similaritas dari ketiga model BERT dibandingkan dengan hasil penilaian persepsional manusia sebagai tolok ukur. Perbandingan digambarkan secara visual menggunakan *Confusion Matrix*. Selanjutnya empat metrik disajikan yaitu accuracy, precision, recall, dan *f1-Score*. Mengingat data untuk tiap kelas

tidak seimbang, nilai *f1-Score* menjadi pertimbangan utama dalam menentukan kinerja algoritma.

Confusion Matrix untuk perhitungan kesesuaian antara teks soal dan CPMK dengan memanfaatkan embedding dari model S-BERT ditunjukan dalam Gambar 5. Gambar tersebut menunjukkan bahwa algoritma dapat memprediksi dengan benar 633 teks soal yang sesuai dengan CPMK, dan dapat memprediksi dengan benar 15 teks soal yang tidak sesuai dengan CPMK.

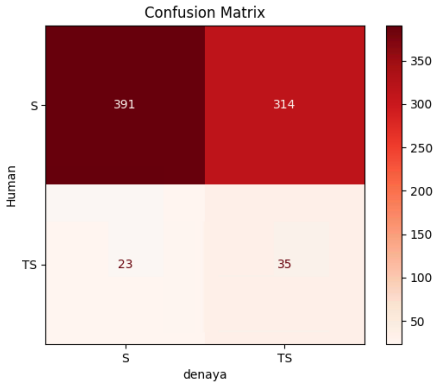


Gambar 5. Confusion Matrix dengan SBERT

Table 4. Analisa model SBERT

	precision	recall	f1-score	support
S	0.94	0.90	0.92	705
TS	0.17	0.26	0.21	58
accuracy			0.85	763
macro avg	0.55	0.58	0.56	763
weight avg	0.88	0.85	0.86	763

Selanjutnya, *Confusion Matrix* untuk perhitungan kesesuaian antara teks soal dan CPMK dengan memanfaatkan embedding dari model Denaya ditunjukan dalam Gambar 6. Gambar memperlihatkan bahwa algoritma hanya mampu memprediksi dengan benar 391 teks soal yang sesuai dengan CPMK, dan memprediksi dengan benar 35 teks soal yang tidak sesuai dengan CPMK.

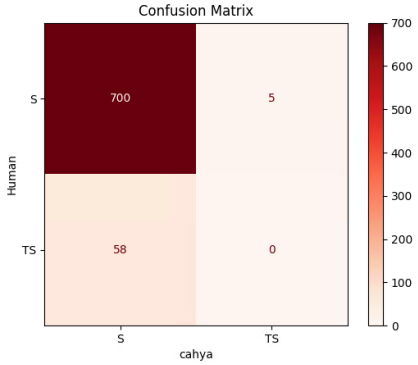


Gambar 6. Confusion Matrix dengan model Denaya

Table 5. Analisa model Denaya

	precision	recall	f1-score	support
S	0.94	0.55	0.70	705
TS	0.10	0.60	0.17	58
accuracy			0.56	763
macro avg	0.52	0.58	0.44	763
weight avg	0.88	0.56	0.66	763

Confusion Matrix untuk perhitungan kesesuaian antara teks soal dan CPMK dengan memanfaatkan embedding dari model Cahya ditampilkan pada Gambar 7. Gambar memperlihatkan bahwa algoritma mampu memprediksi dengan benar 700 teks soal yang sesuai dengan CPMK, namun tidak mampu memprediksi dengan benar teks soal yang tidak sesuai dengan CPMK. Algoritma memberikan prediksi terhadap teks soal yang tidak sesuai CPMK, namun kesemuanya merupakan teks soal yang sesuai CPMK menurut persepsi manusia.



Gambar 7. Confusion Matrix dengan model Cahya

Table 6. Analisa model Cahya

	precision	recall	f1-score	support
S	0.90	0.99	0.96	705
TS	0.00	0.00	0.00	58
accuracy			0.92	763
macro avg	0.46	0.50	0.48	763
weight avg	0.85	0.92	0.88	763

Hasil evaluasi model klasifikasi yang diberikan menunjukkan performa yang beragam. *Model bert-base-nli-mean-tokens* menghasilkan nilai *precision* yang sangat tinggi sebesar 94% dan *recall* 90% serta *f1-score* 92% dengan *accuracy* 85%. Kemudian untuk *model* kedua yaitu *indoSBERT-large* menghasilkan nilai *precision* juga tinggi sebesar 94%, namun nilai *recall* rendah 55% serta *f1-score* 70% dengan nilai *accuracy* 56%. Sedangkan untuk *model* ketiga *bert-base-indonesian-1.5G* menghasilkan nilai *precision* yang cukup tinggi sebesar 90% dan nilai *recall* sangat tinggi 99% serta *f1-score* 96% dengan *accuracy* 92%.

Pembahasan

Perbedaan kinerja antar model deep learning dalam penelitian ini dapat dipahami melalui karakteristik arsitektur, data pelatihan, serta sifat teks pendidikan tinggi yang dianalisis. Tugas mengukur kesesuaian antara rumusan CPMK dan instrumen asesmen tidak hanya terkait pengenalan kemiripan kata, tetapi juga pemahaman semantik terhadap tujuan pembelajaran dan komponen perangkat pembelajaran.

Model SBERT menunjukkan kinerja yang relatif seimbang antara precision dan recall. Hal ini sejalan dengan desain SBERT yang dikembangkan khusus untuk tugas semantic textual similarity melalui arsitektur Siamese dan pembelajaran berbasis kesamaan makna kalimat (Reimers & Gurevych, 2019b). Namun, karena SBERT dilatih pada korpus umum dan multilingual, pemahamannya terhadap bahasa akademik Indonesia dan pernyataan CPMK masih bersifat generik, sehingga beberapa pasangan teks yang secara konseptual tidak sepenuhnya selaras tetap diklasifikasikan sebagai sesuai.

Model Denaya memperlihatkan penurunan kinerja yang besar, terutama pada nilai recall. Temuan ini mengindikasikan bahwa model cenderung gagal mengenali variasi ungkapan konseptual dalam rumusan CPMK dan soal ujian. Salah satu penyebab yang mungkin adalah penggunaan dataset semantic textual similarity hasil terjemahan sebagai data pelatihan, yang berpotensi mengurangi kekayaan representasi semantik bahasa Indonesia alami. Studi sebelumnya menunjukkan bahwa kualitas dan kesesuaian domain data pelatihan sangat berpengaruh terhadap performa model NLP pada konteks spesifik (Nguyen et al., 2019b).

Sebaliknya, model Cahya (bert-base-indonesian-1.5G) menunjukkan kinerja tertinggi pada kelas Sesuai, dengan nilai recall dan akurasi yang sangat tinggi. Keunggulan ini diperdiksi karena proses pretraining menggunakan korpus bahasa Indonesia berskala besar dan beragam, yang memungkinkan model membangun representasi semantik yang lebih kontekstual (Luthfi et al., 2021b). Namun, model ini juga menunjukkan kelemahan yaitu ketidakmampuannya mendeteksi kelas Tidak Sesuai. Kecenderungan untuk mengklasifikasikan hampir seluruh pasangan teks sebagai Sesuai menunjukkan adanya bias terhadap kelas dominan, sebagaimana sering ditemukan pada masalah klasifikasi dengan distribusi data tidak seimbang (Visa et al., 2011).

Dari sisi implikasi praktis, temuan penelitian ini menunjukkan bahwa model deep learning berbasis BERT berpotensi dimanfaatkan sebagai *screening tool* dalam sistem penjaminan mutu dalam implementasi kurikulum. Sistem dapat membantu mengidentifikasi kesesuaian antara CPMK dan instrumen asesmen sebelum dilakukan penilaian manual oleh dosen atau tim penjaminan mutu. Pendekatan ini berpotensi meningkatkan konsistensi, dan objektivitas dalam proses review RPS, sejalan dengan praktik outcome-based education yang menuntut keselarasan konstruktif antara tujuan pembelajaran dan asesmen (D. Pradhan, 2021).

Namun demikian, penggunaan sistem secara otomatis tanpa keterlibatan manusia mengandung risiko yang signifikan. Model berbasis kesamaan semantik belum mampu menangkap aspek pedagogis penting, seperti tingkat kognitif dalam taksonomi Bloom atau kesesuaian strategi asesmen dengan capaian pembelajaran. Selain itu, bias model terhadap kelas dominan dapat menyebabkan asesmen yang sebenarnya tidak selaras tetap lolos proses review. Oleh karena itu, sistem ini sebaiknya diposisikan sebagai pendukung keputusan (*decision support*) dengan pendekatan *human-in-the-loop*, di mana keputusan akhir tetap berada pada dosen. Pendekatan ini sejalan dengan rekomendasi penelitian sebelumnya terkait penerapan kecerdasan buatan dalam konteks pendidikan dan evaluasi akademik (Goularte et al., 2019b; Kuppili et al., 2020b).

Berbagai penelitian telah mengkaji penerapan model deep learning untuk mengukur kesamaan semantik teks. Viji dan Revathy mengusulkan pendekatan hibrida dengan menambahkan komponen *Weighted Fine-Tuner* pada model BERT dan mengombinasikannya dengan Siamese Bi-LSTM, yang diuji pada ratusan ribu pasangan pertanyaan–jawaban dari Quora dan menghasilkan kinerja yang tinggi (Viji & Revathy, 2022). Capaian kinerja tersebut relatif sebanding dengan hasil penelitian ini yang memanfaatkan model BERT yang tidak terlalu kompleks.

Studi lainmenerapkan model BERT untuk menilai kesamaan semantik dokumen klinik berbahasa Jepang, dengan data yang bersumber dari case report dan electronic medical records (Mutinda et al., 2021). Hasilnya adalah bahwa model BERT yang dilatih menggunakan korpus umum berskala besar (Wikipedia) memberikan kinerja yang lebih baik dibandingkan model yang dilatih dengan dokumen klinik spesifik. Temuan ini sejalan dengan hasil penelitian kami, yang menunjukkan bahwa model dengan pretraining bahasa berskala besar mampu menghasilkan akurasi yang sebanding dengan penilaian manusia.

4. Kesimpulan dan Saran

Hasil penelitian menunjukkan bahwa model SBERT dan model Cahya (bert-base-indonesian-1.5G) memiliki potensi untuk digunakan dalam analisis kesesuaian antara rumusan Capaian Pembelajaran Mata Kuliah (CPMK) dan instrumen asesmen pembelajaran. Model Cahya menunjukkan kinerja yang unggul dalam mengidentifikasi pasangan teks yang sesuai, yang tercermin dari nilai recall yang sangat tinggi pada kelas Sesuai. Namun, hasil evaluasi juga menunjukkan bahwa model Cahya belum mampu mendeteksi pasangan teks yang tidak sesuai secara andal, sebagaimana ditunjukkan oleh nilai recall sebesar nol pada kelas Tidak Sesuai.

Dengan demikian, penerapan model Cahya dalam konteks penjaminan mutu kurikulum sebaiknya diposisikan sebagai alat bantu awal (*screening tool*) untuk mengidentifikasi kesesuaian potensial, bukan sebagai sistem penentu keputusan akhir. Keterlibatan penilai manusia tetap diperlukan untuk meninjau kasus-kasus yang memerlukan pertimbangan pedagogis dan akademik yang lebih dalam. Penelitian lanjutan perlu mengkaji strategi untuk meningkatkan kemampuan model dalam mendeteksi ketidaksesuaian, termasuk penyeimbangan data, penyesuaian threshold, dan fine-tuning berbasis domain pendidikan.

Berdasarkan temuan dan keterbatasan tersebut, penelitian lanjutan disarankan untuk menggunakan dataset yang lebih seimbang serta menerapkan pembobotan kesalahan pada kelas Tidak Sesuai agar model lebih sensitif terhadap ketidaksesuaian CPMK dan instrumen asesmen. Selain itu, penyesuaian model menggunakan kumpulan dokumen RPS dan soal ujian perlu dilakukan agar representasi semantik yang dihasilkan lebih relevan dengan konteks pendidikan tinggi.

Ucapan Terimakasih

Penulis mengucapkan terima kasih kepada Direktorat Riset, Teknologi dan Pengabdian Masyarakat (DRTPM), Republik Indonesia yang telah mendanai penelitian ini melalui skema Penelitian Tesis dengan nomor kontrak 006/LL6/PB/AL.04/2023, 170.40/C.1-III/LRI/VI/2023.

Daftar Pustaka

Abas, M. C., & Imam, O. A. (2016). Graduates' Competence on Employability Skills and Job Performance. *International Journal of Evaluation and Research in Education (IJERE)*, *5*(2), 119. https://doi.org/10.11591/ijere.v5i2.4530

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., & Wiebe, J. (2016). {S}em{E}val-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Eds.), *Proceedings of the 10th International Workshop on Semantic Evaluation ({S}em{E}val-2016)* (pp. 497–511). Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1081

Davis, M. H. (2003). *Outcome-Based Education*. *30*(Table 1), 258–263.

Dengel, A., & Dubiel, F. (1995). Clustering and classification of document structure-a machine learning approach. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, *2*, 587–591. https://doi.org/10.1109/ICDAR.1995.601965

Devi, K. U. S., & Suadaa, L. H. (2022). Extractive Text Summarization for Snippet Generation on Indonesian Search Engine using Sentence Transformers. *2022 International Conference on Data Science and Its Applications (ICoDSA)*, 181–186. https://doi.org/10.1109/ICoDSA55874.2022.9862886

Diana, D. (2023). *IndoSBERT: Indonesian SBERT for Semantic Textual Similarity tasks*.

Goularte, F. B., Nassar, S. M., Fileto, R., & Saggion, H. (2019a). A text summarization method based on fuzzy rules and applicable to automated assessment. *Expert Systems with Applications*, *115*, 264–275. https://doi.org/10.1016/j.eswa.2018.07.047

Goularte, F. B., Nassar, S. M., Fileto, R., & Saggion, H. (2019b). A text summarization method based on fuzzy rules and applicable to automated assessment. *Expert Systems with Applications*, *115*, 264–275. https://doi.org/10.1016/j.eswa.2018.07.047

Government of the Republic of Indonesia. (2003). (2003). *Government Regulation of the Republic of Indonesia on the National Education System (Number 20 Year 2003)*.

Kuppili, V., Biswas, M., Edla, D. R., Prasad, K. J. R., & Suri, J. S. (2020a). A Mechanics-Based Similarity Measure for Text Classification in Machine Learning Paradigm. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *4*(2), 180–200. https://doi.org/10.1109/TETCI.2018.2863728

Kuppili, V., Biswas, M., Edla, D. R., Prasad, K. J. R., & Suri, J. S. (2020b). A Mechanics-Based Similarity Measure for Text Classification in Machine Learning Paradigm. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *4*(2), 180–200. https://doi.org/10.1109/TETCI.2018.2863728

Luthfi, E. T., Yusoh, Z. I. M., & Aboobaider, B. M. (2021a). Enhancing the Takhrij Al-Hadith based on Contextual Similarity using BERT Embeddings. *International Journal of Advanced Computer Science and Applications*, *12*(11), 286–293. https://doi.org/10.14569/IJACSA.2021.0121133

Luthfi, E. T., Yusoh, Z. I. M., & Aboobaider, B. M. (2021b). Enhancing the Takhrij Al-Hadith based on Contextual Similarity using BERT Embeddings. *International Journal of Advanced Computer Science and Applications*, *12*(11). https://doi.org/10.14569/IJACSA.2021.0121133

Maharani, S., Nusantara, T., Rahman As'ari, A., & Qohar, A. (2020). *COMPUTITONAL THINKING Pemecahan Masalah di Abad Ke-21*. Wade Group.

Mukherjee, S., Ghosh, M., & Basuchowdhuri, P. (2022). DeepGLSTM: Deep Graph Convolutional Network and LSTM based approach for predicting drug-target binding affinity. *Proceedings of the 2022 SIAM International Conference on Data Mining, SDM 2022*, 729–737. https://doi.org/10.1137/1.9781611977172.82

Mutinda, F. W., Yada, S., Wakamiya, S., & Aramaki, E. (2021). Semantic Textual Similarity in Japanese Clinical Domain Texts Using BERT. *Methods of Information in Medicine*, *60*, E56–E64. https://doi.org/10.1055/s-0041-1731390

Nguyen, H. T., Duong, P. H., & Cambria, E. (2019a). Learning short-text semantic similarity with word embeddings and external knowledge sources. *Knowledge-Based Systems*, *182*, 104842. https://doi.org/10.1016/j.knosys.2019.07.013

Nguyen, H. T., Duong, P. H., & Cambria, E. (2019b). Learning short-text semantic similarity with word embeddings and external knowledge sources. *Knowledge-Based Systems*, *182*, 104842. https://doi.org/10.1016/j.knosys.2019.07.013

Nurjannah, I., Cholik, M., Theodorus, M., Vinaya, D., & Made, W. I. (2021). *Development of OBE-Based Learning Evaluation Model in Mechanical Engineering Education Program*. *209*(Ijcse), 7–12.

Pradhan, Devasis. (2021). Effectiveness of Outcome Based Education (OBE) toward Empowering the Students Performance in an Engineering Course. *Journal of Advances in Education and Philosophy*, *5*(2), 58–65. https://doi.org/10.36348/jaep.2021.v05i02.003

Pradhan, D. (2021). Effectiveness of Outcome Based Education (OBE) toward Empowering the Students Performance in an Engineering Course. *Journal of Advances in Education and Philosophy*, *5*(2), 58–65. https://doi.org/10.36348/jaep.2021.v05i02.003

Reimers, N., & Gurevych, I. (2019a). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. https://doi.org/10.18653/v1/d19-1410

Reimers, N., & Gurevych, I. (2019b). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3980–3990. https://doi.org/10.18653/v1/D19-1410

Rojak, A., Wasliman, I., Koswara, N., & Karim Fatkhullah, F. (2022). Learning Management In Increasing The Quality Of Graduates At Aliyah Boarding Schools In Banten Province (Study on MAN Insan Cendekia and MAN 2 Serang). *International Journal of Educational Research & Social Sciences*, *3*(5), 2166–2177. https://doi.org/10.51601/ijersc.v3i5.524

Subadi, T., Narimo, S., & Hidayati, E. F. (2022). Based Learning Training Lesson Study to Improve the Quality of Elementary School Teachers Kartasura Muhammadiyah, Indonesia. *Warta LPM*, *25*(1), 1–9. https://doi.org/10.23917/warta.v25i1.592

T S, A., P C, R., & Murali, R. (2019). Text to SQL Query Conversion Using Deep Learning: A Comparative Analysis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3437616

Thamrin, H., & Sabardila, A. (2016). Utilizing Lexical Relationship in Term-Based Similarity Measure Improves Indonesian Short Text Classification. *ARPN Journal of Engineering and Applied Sciences*, *11*(22), 13141–13145.

Thamrin, H., Verdikha, N. A., & Triyono, A. (2021). Text Classification and Similarity Algorithms in Essay Grading. *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 201–205.

Viji, D., & Revathy, S. (2022). A hybrid approach of Weighted Fine-Tuned BERT extraction with deep Siamese Bi – LSTM model for semantic text similarity identification. *Multimedia Tools and Applications*, *81*(5), 6131–6157. https://doi.org/10.1007/s11042-021-11771-6

Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *Maics*, *710*(1), 120–127.

Zade, N., & Ajani, S. (2022). Multilingual text classification using deep learning. *International Journal of Health Sciences*, *11*(06), 10528–10536. https://doi.org/10.53730/ijhs.v6ns1.7539