Jurnal Teknik Indonesia

Volume 5 Nomor 3 July 2026

https://jti.publicascientificsolution.com/index.php/rp

Pengembangan Model NLP Untuk Penilaian Jawaban Esai Otomatis Menggunakan SBERT dan Stanza

Riva Iqbal Almero
Universitas Widyatama, Indonesia
Email: riva.almero@widyatama.ac.id

Keywords:

Automated Essay Scoring, SBERT, Stanza, Multi-Layer Perceptron, Semantic Bias.

Abstract

Manual essay assessment processes face several limitations, including evaluator subjectivity, lengthy evaluation time, and inconsistency in score assignment. The advancement of Natural Language Processing (NLP) provides opportunities to develop Automated Essay Scoring (AES) systems capable of understanding semantic and linguistic characteristics of student responses. This study aims to develop an NLP-based model for automatic essay response relevance classification by combining Sentence-Bidirectional Encoder Representations from Transformers (SBERT) and Stanza. This research applies a quantitative approach using a Research and Development (R&D) method. The dataset was obtained from the UKARA 1.0 Challenge, consisting of 2,861 Indonesian short essay responses. The research stages include text preprocessing, semantic feature extraction using SBERT, linguistic feature extraction using Stanza, mechanical feature extraction based on spelling errors, and model development using the Multi-Layer Perceptron (MLP) algorithm. The evaluation results demonstrate that the proposed model achieved an accuracy of 0.66 and an F1-score of 0.76 for the relevant response class. However, the model showed limitations in identifying irrelevant responses, achieving only 0.34 recall for the negative class due to the dominance of semantic features. Hyperparameter experiments revealed that architectural modifications did not significantly improve model performance. In conclusion, the combination of SBERT and Stanza provides a promising approach for automated essay scoring systems; however, further improvements are required through semantic bias mitigation strategies and additional feature development to enhance classification performance.

Kata Kunci

Penilaian Esai Otomatis, SBERT, Stanza, Multi-Layer Perceptron, Bias Semantik

Abstrak

Proses penilaian jawaban esai secara manual memiliki berbagai keterbatasan, seperti subjektivitas penilai, waktu evaluasi yang lama, serta ketidakkonsistenan dalam pemberian skor. Perkembangan Natural Language Processing (NLP) memungkinkan pengembangan sistem penilaian esai otomatis (Automated Essay Scoring) yang mampu memahami karakteristik semantik dan linguistik dari suatu jawaban. Penelitian ini bertujuan untuk mengembangkan model NLP dalam melakukan klasifikasi relevansi jawaban esai otomatis menggunakan kombinasi Sentence-Bidirectional Encoder Representations from Transformers (SBERT) dan Stanza. Metode penelitian menggunakan pendekatan kuantitatif dengan metode research and development (R&D). Dataset yang digunakan berasal dari UKARA 1.0 Challenge dengan jumlah 2.861 data jawaban esai berbahasa Indonesia. Tahapan penelitian meliputi preprocessing, ekstraksi fitur semantik menggunakan SBERT, ekstraksi fitur linguistik menggunakan Stanza, ekstraksi fitur mekanik berdasarkan kesalahan ejaan, serta pembangunan model menggunakan algoritma Multi-Layer Perceptron (MLP). Hasil evaluasi menunjukkan

bahwa model mampu mencapai akurasi sebesar 0,66 dengan nilai F1-score kelas relevan sebesar 0,76. Namun, model mengalami keterbatasan dalam mengenali jawaban tidak relevan dengan nilai recall kelas negatif sebesar 0,34 akibat dominasi fitur semantik. Eksperimen variasi hyperparameter menunjukkan bahwa perubahan arsitektur tidak memberikan peningkatan signifikan terhadap performa model. Kesimpulannya, kombinasi SBERT dan Stanza mampu mendukung sistem penilaian esai otomatis, tetapi diperlukan strategi mitigasi bias semantik dan pengembangan fitur tambahan untuk meningkatkan kemampuan klasifikasi

PENDAHULUAN

Evaluasi hasil belajar dilakukan untuk mengetahui sejauh mana peserta didik mampu memahami materi yang telah dipelajari selama proses pembelajaran. Evaluasi hasil belajar dapat dilakukan dalam banyak bentuk, salah satunya adalah melalui tes subjektif berbentuk esai. Berbeda dengan pilihan ganda, ujian esai dapat mengukur kemampuan peserta didik terhadap pengetahuan yang telah diperolehnya (Setio Pribadi et al., 2023). Proses penilaian esai secara manual memiliki beberapa tantangan, seperti subjektivitas penilai, waktu yang dibutuhkan untuk mengevaluasi jawaban, serta konsistensi dalam pemberian nilai.

Seiring dengan perkembangan *Artificial Intelligence* (AI), khususnya di bidang *Natural Language Processing* (NLP), telah muncul berbagai penelitian untuk membuat sistem penilaian esai otomatis yang memanfaatkan NLP. NLP memungkinkan komputer untuk menganalisis, memahami, dan menghasilkan bahasa manusia, termasuk ujaran yang biasa diucapkan oleh manusia (Purnamasari et al., 2023). Penggunaan NLP sudah diterapkan dalam berbagai bidang dalam kehidupan sehari-hari, termasuk dalam bidang pendidikan, seperti peringkasan teks dan *paraphrasing*, tanya-jawab, *chatbot*, dan evaluasi ejaan dan *grammar* (Rumaisa et al., 2021).

Beberapa penelitian telah dilakukan mengenai pengembangan sistem penilaian otomatis berbasis NLP, seperti yang dilakukan oleh Chamidah & Santoni (2021) yang berjudul "Pencocokan Berbasis Kata Kunci dan Penilaian Esai Pendek Otomatis Berbahasa Indonesia". Penelitian tersebut melakukan penilaian esai dengan menghitung kecocokan antara jawaban uji dengan jawaban referensi dengan melihat kata kunci dari masing-masing jawaban. Hasil dari penelitian menunjukkan skor *Mean Absolute Error* untuk keseluruhan jawaban uji sebesar 0.25 dan korelasi antara nilai uji dan nilai *grading* sebesar 79%.

Cahyadi et al. (2025) pada penelitiannya yang berjudul "Penilaian Esai Mata Kuliah Bahasa Inggris Mahasiswa Berbasis *Machine Learning* Menggunakan Algoritma Regresi Linier" mengusulkan model berbasis algoritma regresi linier untuk menilai esai mahasiswa berdasarkan fitur linguistik dan struktural. Model ini mengidentifikasi berbagai fitur teks seperti jumlah kata, panjang kalimat, kompleksitas kata, serta pola gramatikal yang berkontribusi dalam penentuan skor akhir esai. Penerapan sistem ini dapat membantu pengajar dalam memberikan umpan balik yang lebih objektif dan efisien. Namun terdapat beberapa keterbatasan, seperti kurangnya kemampuan dalam memahami konteks secara mendalam.

Salah satu pendekatan yang dapat dilakukan untuk menangani penilaian semantik adalah penggunaan *Sentence-Bidirectional Encoder Representation from Transformer* (SBERT). SBERT mengintegrasikan *Siamese Network* dengan model BERT yang telah dilatih untuk menghasilkan *sentence embedding* yang bermakna secara semantik. SBERT dapat dikombinasikan dengan Stanza, sebuah perangkat pemrosesan bahasa alami yang dapat mengolah struktur bahasa seperti *grammar* yang memiliki tingkat akurasi yang tinggi (Qi et al., 2020).

Penelitian ini bertujuan untuk mengeksplorasi opsi model sistem penilaian esai berbahasa Indonesia menggunakan SBERT yang dikombinasikan dengan Stanza. Pendekatan ini

diharapkan dapat menghasilkan model penilaian yang tidak hanya mampu menilai tingkat kesamaan suatu kalimat, tetapi juga mampu memahami konteks dan sensitif terhadap kualitas linguistik.

Natural Language Processing (NLP) merupakan salah satu cabang dari Artificial Intelligence (AI) yang berfokus pada pengembangan sistem yang mampu memahami bahasa manusia secara alami (Miharja & Adhkar, 2022). Lebih lanjut, NLP dapat dipahami sebagai teknik pemrograman yang memungkinkan komputer memahami dan menghasilkan output dalam bahasa manusia (F. P. E. Putra & Asmunin, 2024). Salah satu penerapan NLP yang relevan dalam konteks penelitian ini adalah Automated Essay Scoring (AES), yakni sebuah sistem evaluasi berbasis tugas regresi teks yang bertujuan menilai esai secara otomatis dengan memberikan skor sesuai kualitas penulisannya (Yang et al., 2020). Sistem ini bekerja dengan membandingkan jawaban referensi yang telah ditentukan dengan jawaban yang dinilai sehingga mempermudah proses pemeriksaan jawaban esai (Ahmad & Sasue, 2020).

Dalam proses pemrosesan teks, penelitian ini memanfaatkan beberapa metode dan alat bantu. Pertama, Sentence-Bidirectional Encoder Representation from Transformer (SBERT), yaitu metode yang menghitung representasi vektor tingkat kalimat menggunakan arsitektur dua jaringan BERT berbobot sama, di mana masing-masing jaringan memproses satu kalimat masukan (Chi et al., 2023). SBERT dilatih menggunakan arsitektur Siamese Network dan Triplet Network guna menghasilkan sentence embedding yang bermakna secara semantik, serta menambahkan operasi pooling pada output BERT untuk membentuk representasi berukuran tetap dari masukan berupa frasa yang panjangnya bervariasi (Reimers & Gurevych, 2019; Santander-Cruz et al., 2022). Kedua, Stanza, yaitu perangkat pemrosesan bahasa alami berbasis Python yang mendukung banyak bahasa manusia. Stanza memiliki pipeline berbasis neural yang mengambil teks mentah sebagai masukan dan menghasilkan berbagai anotasi linguistik, meliputi tokenisasi, lemmatisasi, penandaan part-of-speech, dependency parsing, hingga named entity recognition (Qi et al., 2020). Dependency parser pada Stanza merupakan parser neural deep biaffine berbasis Bi-LSTM yang ditingkatkan dengan fitur tambahan bermotivasi linguistik, sehingga menghasilkan performa yang lebih tinggi dan konsisten meskipun meningkatkan kompleksitas dan waktu proses (Zupon et al., 2022).

Untuk mengukur tingkat kesamaan antar teks, digunakan Cosine Similarity, yakni algoritma yang mengukur kesamaan dua dokumen berdasarkan sudut kosinus antara dua vektor yang diproyeksikan dalam ruang multidimensi (Rinjeni et al., 2024). Adapun tingkat kesamaan tersebut dinyatakan dengan notasi matematis sebagai berikut:

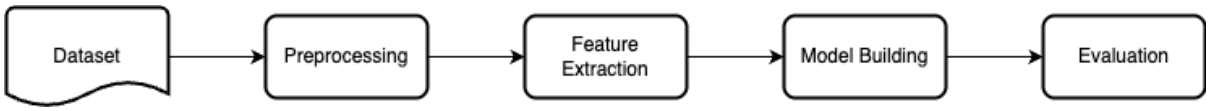
$$\cos\theta = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Selanjutnya, proses pembangunan model dalam penelitian ini menggunakan Keras, yaitu Application Programming Interface (API) berbasis TensorFlow yang dirancang untuk mempermudah perancangan, pelatihan, serta evaluasi model jaringan neural dan dikenal karena fleksibilitasnya dalam menerapkan beragam algoritma deep learning (Chicho & Bibo Sallow, 2021; Ouahi et al., 2024). Dua algoritma utama yang diterapkan adalah Multi-Layer Perceptron (MLP) dan Random Forest Classifier. MLP merupakan varian feed-forward neural network yang terdiri atas beberapa layer tersembunyi berisi neuron-neuron, dan mampu merepresentasikan hubungan yang kompleks dan nonlinier serta menunjukkan kinerja yang baik dalam tugas prediksi, estimasi fungsi, dan pengenalan pola (Al Bataineh et al., 2022; Alla et al., 2021). Sementara itu, Random Forest Classifier merupakan algoritma machine learning yang memanfaatkan kumpulan pohon keputusan untuk melakukan klasifikasi data secara efisien, di mana setiap pohon menghasilkan prediksi dari sampel data yang dipilih secara acak

dan hasil akhir ditentukan berdasarkan voting mayoritas (Pradana et al., 2024; Pangkasidhi et al., 2022).

METODE

Penelitian ini bersifat kuantitatif menggunakan pendekatan *research and development* (R&D) dengan fokus pada pengembangan model NLP untuk sistem penilaian otomatis. Model yang dikembangkan memanfaatkan kombinasi dari SBERT dan Stanza. Penelitian ini bersifat ekperimental, karena model akan dibangun, dilatih, dan dievaluasi secara langsung menggunakan data yang dikumpulkan. Diagram metodologi penelitian disajikan pada Gambar 1.



Gambar 1 Diagram Metodologi Penelitian

Dataset yang digunakan bersumber dari UKARA 1.0 Challenge (I. F. Putra, 2019) berjumlah 2.861 entri yang berupa jawaban esai singkat dari 2 stimulus yang berbeda, serta label di setiap entri. Label bersifat biner dimana 1 merepresentasikan jawaban yang relevan dan 0 merepresentasikan jawaban yang tidak relevan. Contoh dataset disajikan pada Tabel 1.

Tabel 1 Contoh Dataset

RES_ID	RESPONSE	LABEL
TRA1	intetraksi/beradaptasi terhadap lingkungan yang baru.	1
TRA2	seperti jatuhnya meteor tsunami gempa bumi	0
TRA3	hanya tuhan yang tahu tantangan nya itu apaan	0
TRA4	mereka akan sulit beradaptasi	1
TRA5	Tempat tinggal, ekonomi, dan pekerjaan	1

Dataset terdiri dari 3 tipe set yang berbeda: (1) *Training Set* untuk melatih model; (2) *Development Set* untuk memantau kinerja model selama pelatihan; dan (3) *Test Set* untuk evaluasi akhir model. Pembagian dataset mengikuti skenario pembagian baku dari sumber dataset. Detail tipe set dataset disajikan pada Tabel 2.

Tabel 2 Detail Tipe Set Dataset

Tipe Set	Data A	Data B
Training	268	305
Development	215	244
Test	855	974

Selain itu, terdapat dataset tambahan untuk membantu proses penelitian. Dataset tambahan yang digunakan yaitu daftar kata yang diambil dari Kamus Besar Bahasa Indonesia. Dataset daftar kata digunakan dalam proses pengecekan ejaan pada proses ekstraksi fitur.

Preprocessing

Preprocessing dilakukan untuk mendapatkan kumpulan data akhir yang berguna untuk proses selanjutnya. Data yang bersih penting untuk memastikan model dapat fokus pada makna semantik dan tidak terdistorsi oleh *noise*. *Preprocessing* dilakukan seminimal mungkin dengan melakukan HTML *decoding*, normalisasi *whitespace*, dan menghilangkan karakter non-alfanumerik yang bukan tanda baca. *Stemming* dan *stopword removal* tidak dilakukan untuk menjaga keutuhan struktur kalimat yang dibutuhkan oleh SBERT dan Stanza.

Ekstraksi Fitur

Fitur yang digunakan meliputi: (1) fitur relevansi (SBERT); (2) fitur linguistik (Stanza); serta (3) fitur mekanik (rasio kesalahan ejaan). Tiap fitur diekstrak menggunakan metode yang berbeda. Fitur relevansi diekstrak menggunakan SBERT varian "*paraphrase-multilingual-mpnet-base-v2*" yang mendukung bahasa Indonesia. Proses dilakukan dengan melakukan *encoding* pada jawaban pendek esai dan simulus untuk mendapatkan vektor dari teks, kemudian kemiripannya dihitung menggunakan *Cosine Similarity*. Hasil yang didapatkan digunakan sebagai fitur relevansi. Fitur linguistik diekstrak menggunakan Stanza untuk menghasilkan nilai-nilai total kata, total kalimat, rata-rata panjang kalimat, rasio kata unik, dan rasio *part-of-speech* terhadap total kata. Fitur mekanik diekstrak dengan cara membandingkan setiap kata yang ada di jawaban esai dengan Daftar Kata Bahasa Indonesia untuk mendapatkan rasio kesalahan ejaan terhadap jumlah kata.

Arsitektur Model

Penelitian ini menerapkan *artificial neural network* jenis *Multi-Layer Perceptron* (MLP) untuk mengklasifikasikan jawaban esai ke dalam kategori "Relevan" (1) atau "Tidak Relevan" (0). MLP dipilih karena kemampuannya dalam memodelkan hubungan non-linear yang kompleks antara fitur input dengan label target.

Secara struktural, model yang dibangun terdiri dari komponen-komponen sebagai berikut:

a. *Input Layer*
   Layer ini menerima vektor fitur yang telah melalui proses normalisasi menggunakan teknik *Standard Scaling*. Normalisasi dilakukan untuk mempercepat proses pelatihan model dan mencegah dominasi fitur dengan rentang nilai besar terhadap fitur lainnya.
b. *Hidden Layer*
   Model ini menggunakan arsitektur MLP dengan beberapa lapisan tersembunyi yang disusun secara menurun untuk mengekstraksi fitur abstrak secara bertahap. Arsitektur spesifik yang digunakan terdiri dari dua h*idden layer* yang secara berurutan terdiri dari 64 dan 32 neuron. Setiap neuron pada *hidden layer* menggunakan fungsi aktivasi ReLU (*Rectified Linear Unit*). Untuk mencegah *overfitting*, teknik regularisasi *Dropout* diterapkan setelah setiap *hidden layer* dengan *rate* sebesar 0.3. *Dropout* bekerja dengan cara menonaktifkan neuron secara acak selama proses pelatihan, sehingga model dipaksa untuk mempelajari fitur yang lebih *robust* dan tidak bergantung pada neuron tertentu.
c. *Output Layer*
   Output layer terdiri dari satu neuron tunggal yang menggunakan fungsi aktivasi Sigmoid. Fungsi ini mentransformasikan hasil pembobotan akhir menjadi nilai antara 0 dan 1.

Selain itu, model juga dilengkapi dengan Adam *optimizer* dengan *learning rate* rendah agar proses *training* model berjalan lebih stabil. Adam digunakan sebagai *optimizer* karena kemampuannya dalam mengatur *learning rate* secara mandiri untuk setiap parameter, sehingga *learning rate* dapat dikonfigurasi dengan lebih fleksibel.

Skenario Pengujian dan Evaluasi

Tahap pengujian dan evaluasi meliputi memprediksi *Test Set* menggunakan model yang sudah dibuat, kemudian menganalisa *false positive* dan *false negative*, visualisasi fitur yang paling penting, dan evaluasi hasil prediksi model. Analisa *false positive* dan *false negative* dilakukan dengan membandingkan hasil prediksi model dengan label asli dari dataset. Visualisasi fitur yang paling penting dilakukan dengan menggunakan *Random Forest Classifier* dan memvisualisasikannya menjadi diagram. Hasil prediksi model kemudian dievaluasi dengan melihat nilai *precision, recall,* dan *f1-score* dari hasil prediksi.

HASIL DAN PEMBAHASAN

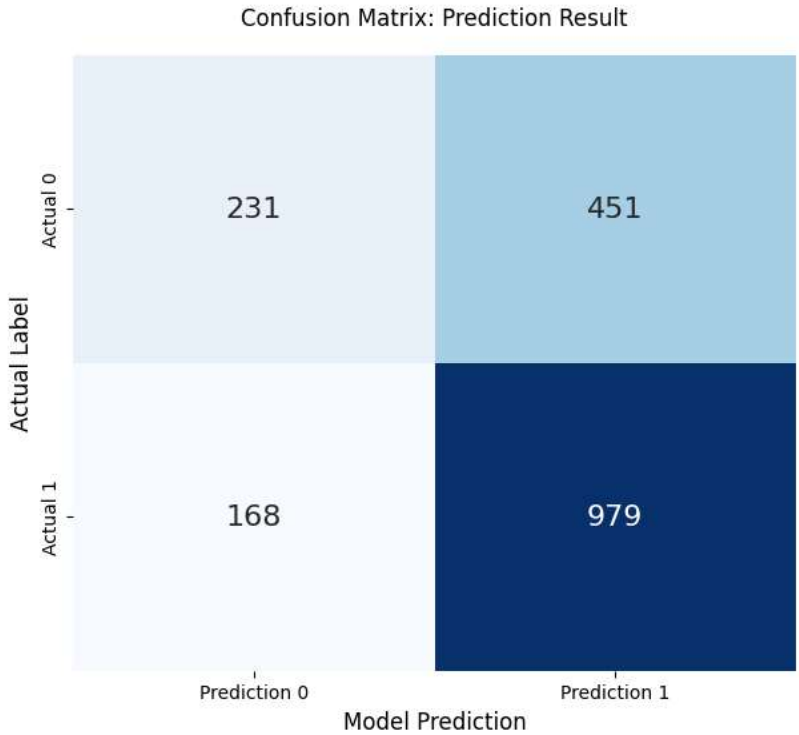
Pembuatan model dilakukan dengan melakukan *preprocessing* terhadap dataset. Kemudian dilakukan ekstraksi fitur relevansi menggunakan SBERT, fitur linguistik menggunakan Stanza, dan fitur mekanik dengan bantuan dataset daftar kata Bahasa Indonesia. Selanjutnya model dibangun menggunakan Keras. Model dilatih menggunakan *Train Set* dan divalidasi menggunakan *Development Set*. Terakhir, model dievaluasi menggunakan *Test Set* yang belum pernah model lihat sebelumnya.

Tabel 3 Hasil Evaluasi Prediksi Test Set

	precision	recall	f1-score	support
0	0.58	0.34	0.43	682
1	0.68	0.85	0.76	1147
accuracy			0.66	1829
macro avg	0.63	0.60	0.59	1829
weighted avg	0.65	0.66	0.64	1829

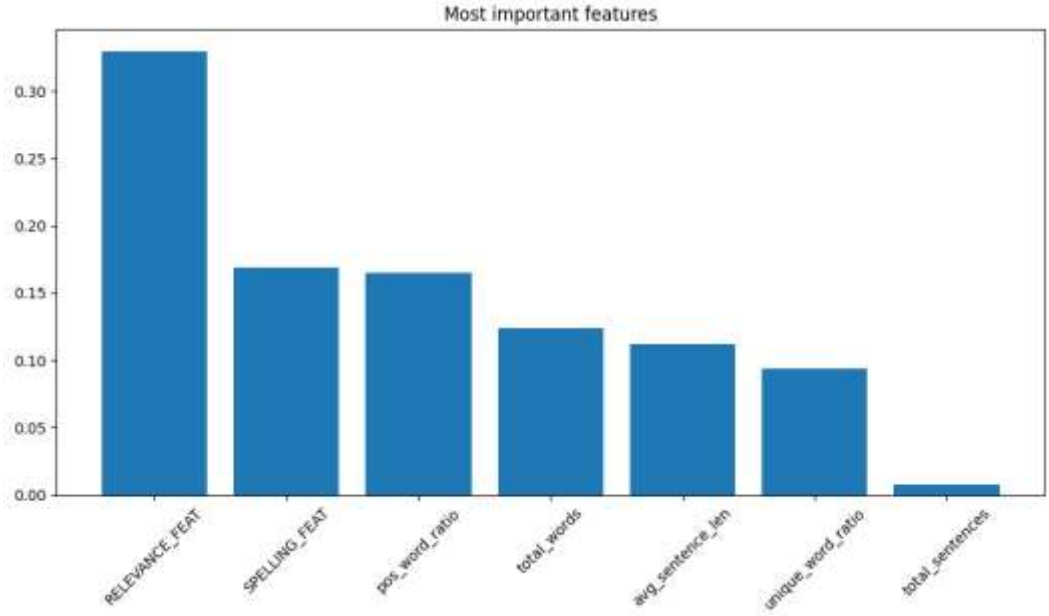
Berdasarkan *Classification Report* dari *Test Set* gabungan yang berjumlah 1829 data, performa model kelas positif (Label 1) memperlihatkan *f1-score* sebesar 0.76. Model menunjukkan kemampuan yang cukup dalam mengidentifikasi jawaban yang relevan, terlihat dari nilai kelas positif yang mencapai 0.85. Namun, terdapat keterbatasan yang cukup signifikan dalam mendeteksi jawaban tidak relevan, di mana *recall* kelas negatif hanya mencapai nilai 0.34. Hal ini mengungkap temuan bahwa model gagal mendeteksi hampir 70% jawaban yang salah.

Selain itu, *Classification Report* juga menunjukkan nilai *macro avg* sebesar 0.59 dan *weighted avg* sebesar 0.64. Hal ini mengindikasikan bahwa model mengalami kesulitan dalam memprediksi kelas minoritas secara akurat. Dalam konteks penelitian ini, rendahnya nilai *macro avg* mencerminkan keterbatasan model dalam menangani distribusi data yang bersifat *imbalanced.*



Gambar 2 Visualisasi *Confusion Matrix* dari Hasil Prediksi Model

Rendahnya nilai *recall* kelas negatif berkaitan dengan tingginya jumlah kasus *false positive*. Model menghasilkan prediksi yang salah pada kelas negatif sebanyak 451 data, yang menjadi kasus *false positive*. Hal ini mengindikasikan bahwa model belum mampu membedakan secara efektif antara jawaban yang memiliki makna konten yang benar dengan jawaban yang hanya mengandung kata kunci relevan tanpa kesesuaian konteks yang memadai.

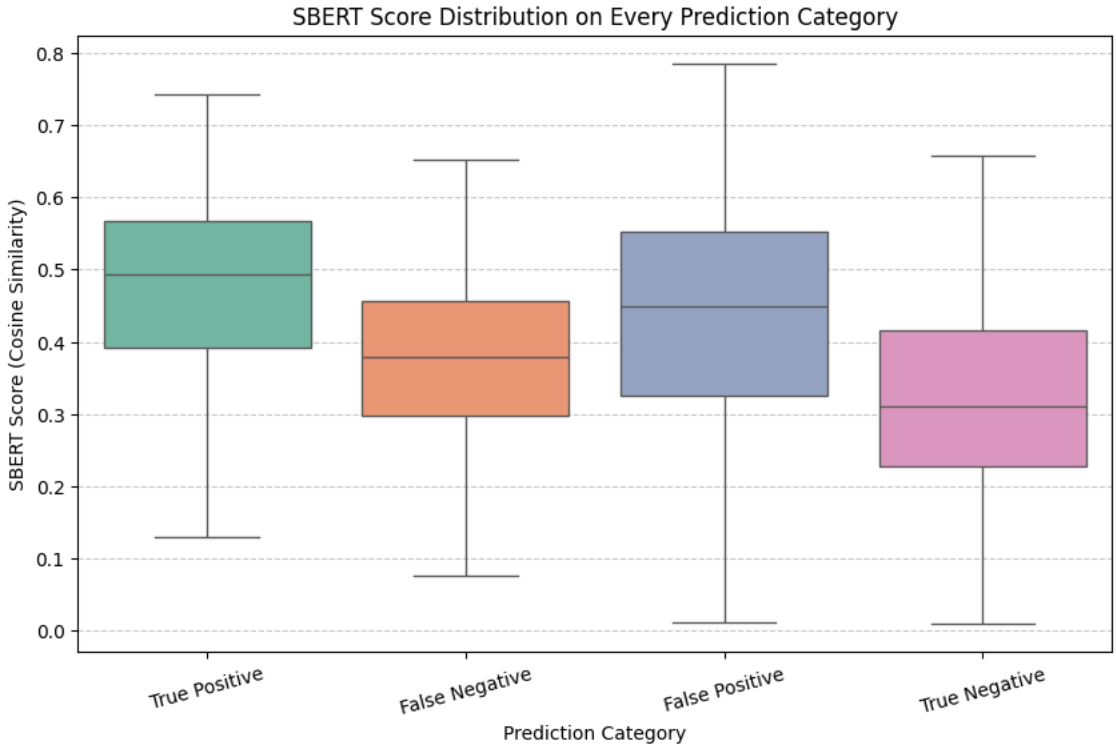


Gambar 3 Visualisasi Kontribusi Fitur

Berdasarkan visualisasi kontribusi fitur, terdapat temuan bahwa fitur relevansi semantik menunjukkan dominasi yang sangat signifikan dengan bobot kontribusi lebih dari 30%. Hal ini mengindikasikan bahwa model MLP yang dibangun sangat bergantung pada kemiripan vektor antara jawaban dengan stimulus. Dominasi ini menjadi alasan utama tingginya angka *false positive*, di mana model cenderung memberikan prediksi positif hanya karena adanya kesamaan kata kunci, meskipun padanan kalimat yang digunakan mungkin tidak tepat.

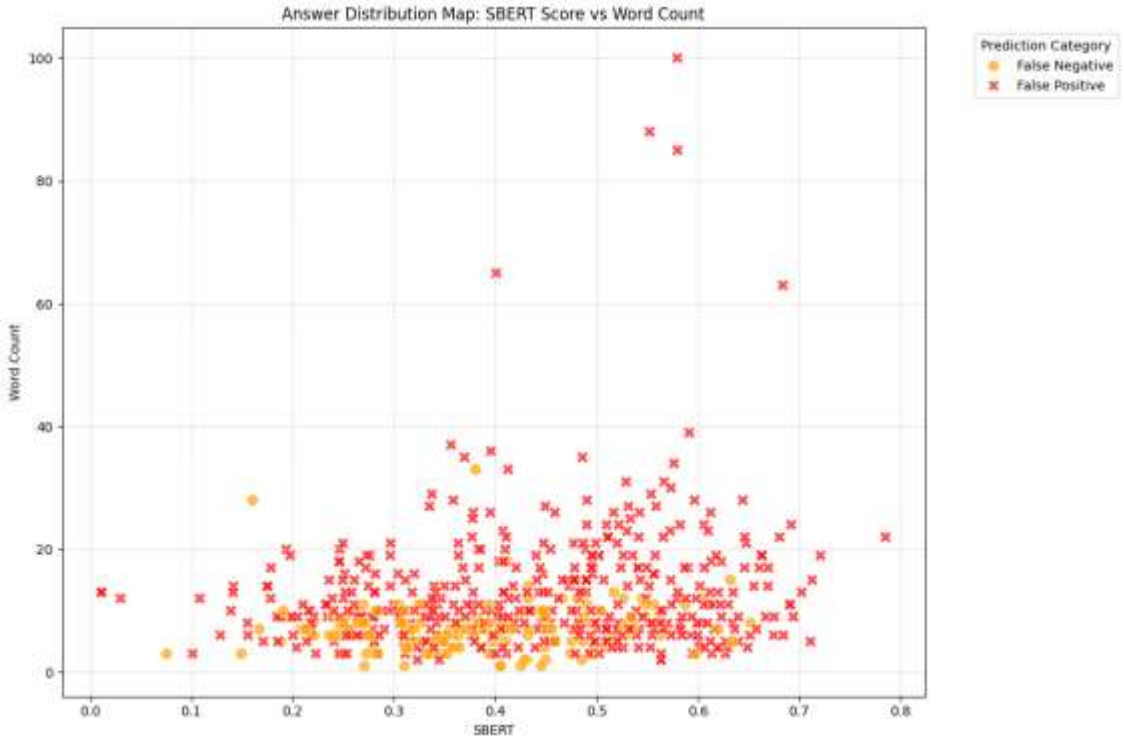
Fitur mekanik dan fitur-fitur linguistik lainnya hanya berkontribusi tidak lebih dari 20%, yang berarti peran kualitas ejaan dan struktur tata bahasa belum mampu mengimbangi kuatnya fitur semantik dari SBERT dalam proses pengambilan keputusan model. Fitur mekanik memiliki nilai bobot sebesar ~17%, yang menunjukkan bahwa model memberikan atensi yang cukup tinggi terhadap penulisan ejaan sebagai indikator kualitas jawaban. Namun, fitur mekanik belum mampu mengimbangi dominasi fitur semantik. Fitur linguistik sebagian besar berkontribusi di bawah 13% dengan angka kumulatif mencapai ~50%. Namun hal ini tidak mengubah fakta bahwa model lebih condong terhadap fitur semantik.

Model yang dibangun menggunakan MLP dirancang untuk memodelkan hubungan non-linear. Dalam konteks penilaian yang berdasar pada relevansi, SBERT secara alami membawa informasi yang lebih kuat dalam hal kesamaan konteks. Karena itulah model menunjukkan ketergantungan yang lebih tinggi pada fitur semantik dibanding fitur-fitur lainnya.



Gambar 4 Visualisasi *Boxplot* Terhadap Distribusi Skor SBERT pada Tiap Kategori Prediksi

Visualisasi *Box Plot* menjadi bukti atas rendahnya skor *recall* negatif yang telah dibahas sebelumnya. Terlihat adanya tumpang tindih yang ekstrem antara distribusi skor SBERT pada kategori *true positive* dan *false positive*, dengan nilai median yang hampir identik di kisaran 0.45 – 0.49. Ketidakmampuan SBERT dalam membedakan kedua kelas ini secara otomatis membuat model MLP gagal melakukan klasifikasi yang presisi. Model menjadi sangat sensitif terhadap kehadiran kata kunci yang relevan namun gagal melakukan verifikasi konteks, sehingga jawaban salah dengan *surface-level similarity* seringkali diklasifikasikan sebagai jawaban benar.



Gambar 5 Visualisasi *Scatter Plot* Perbandingan Jumlah Kata dengan Skor SBERT pada kasus *False Positive* dan *False Negative*

Analisa lebih dalam mengenai *Scatter Plot* mengungkap fenomena *surface-levevl similarity*. Kesalahan tipe *false positive* yang mendominasi rentang skor SBERT menengah ke atas menunjukkan bahwa model rentan terhadap manipulasi jawaban berupa repetisi kata kunci atau jawaban panjang yang tidak substansial. Di sisi lain, persebaran *false negative* pada nilai SBERT di bawah 0.45 mengonfirmasi bahwa model gagal mengenali jawaban akurat yang diekspresikan secara singkat atau melalui parafrasa yang berbeda dari stimulus. Hal ini memperkuat temuan bahwa ketergantungan berlebih pada skor semantik mengakibatkan model mengabaikan kualitas logika dan struktur kalimat yang seharusnya disediakan oleh fitur linguistik.

Tabel 4 Konfigurasi Model untuk Eksplorasi Hyperparameter

Konfigurasi	Neuron Layer 1	Neuron Layer 2	Dropout Rate	Learning Rate
Baseline	64	32	0.3	0.0001
Higher LR	64	32	0.3	0.001
Wider	128	64	0.4	0.0001
Deep & Small	32	16	0.2	0.0005

Analisis terhadap *hyperoarameter* dilakukan untuk memvalidasi apakah arsitektur model awal sudah merupakan konfigurasi paling optimal. Eksperimen dilakukan dengan memodifikasi jumlah neuron pada NLP, *learning rate*, dan *dropout rate*.

Tabel 5. Perbandingan Kinerja Model MLP pada Berbagai Konfigurasi Hyperparameter

Konfigurasi	Accuracy	Recall (0)	Recall (1)	F1-Score (0)	F1-Score (1)
Baseline	0.66	0.37	0.83	0.45	0.75
Higher RL	0.66	0.37	0.84	0.45	0.76
Wider	0.66	0.37	0.83	0.45	0.76
Deep & Small	0.66	0.37	0.83	0.45	0.76

Temuan paling krusial dari eksplorasi ini adalah stagnansi nilai akurasi pada angka 0.66 dan *recall* kelas negatif pada angka 0.37 di seluruh konfigurasi. Meskipun kapasitas model ditingkatkan maupun disederhanakan, model tidak menunjukkan peningkatan sensitivitas yang signifikan dalam mendeteksi jawaban yang tidak relevan. Hal ini membuktikan bahwa limitasi sistem penilaian ini tidak terletak pada kedalaman NLP, melainkan pada kualitas fitur *input* itu sendiri.

Eksperimen ini menyimpulkan bahwa model *Baseline* maupun *Deep & Small* sudah merupakan representasi optimal dari dataset yang tersedia. Ketidakmampuan model untuk melewati ambang akurasi 0.66 meskipun sudah dilakukan *tuning* arsitektur mengonfirmasi bahwa arah pengembangan selanjutnya harus pada mitigasi bias semantik pada tingkat ekstraksi fitur.

KESIMPULAN

Implementasi SBERT dan Stanza dalam sistem penilaian esai otomatis pada penelitian ini mengungkap tantangan utama dalam penanganan bias semantik. Meskipun model mampu mengenali jawaban benar dengan cukup akurat dengan perolehan *f1-score* sebesar 0.76, evaluasi mendalam menunjukkan bahwa ketergantungan model yang berlebihan pada skor semantik menurunkan sensitivitas terhadap fitur linguistik dan mekanik. Temuan ini dibuktikan dengan skor *recall* kelas negatif yang hanya sebesar 0.34 dan kontribusi fitur semantik yang mencapai lebih dari 30%. Hal ini mengakibatkan tingginya angka *false positive*, di mana model gagal membedakan jawaban yang berkualitas dengan jawaban yang hanya memiliki kata kunci yang beririsan.

Hasil penelitian ini menunjukkan bahwa penggunaan *sentence embedding* saja belum cukup untuk menangani variasi jawaban pendek. Oleh karena itu, beberapa hal yang dapat dilakukan untuk pengembangan penelitian ini antara lain melakukan mitigasi bias semantik untuk mengurangi kasus *false positive*, menerapkan aturan sederhana berupa ambang batas jumlah kata minimum sebagai bentuk mitigasi kelemahan model dalam memprediksi jawaban pendek tapi memiliki kata kunci yang relevan, serta menerapkan *fine-tuning* pada model SBERT secara langsung.

REFERENSI

Ahmad, R., & Sasue, R. R. O. (2020). Sistem Penilaian Esai Otomatis Menggunakan Algoritma Stemming Nazief dan Adriani. *Jurnal Teknologi Transportasi Dan Logistik*, *1*(2), 101–108.

Al Bataineh, A., Kaur, D., & Jalali, S. M. J. (2022). Multi-Layer Perceptron Training Optimization Using Nature Inspired Computing. *IEEE Access*, *10*, 36963–36977. https://doi.org/10.1109/ACCESS.2022.3164669

Alla, H., Moumoun, L., & Balouki, Y. (2021). A Multilayer Perceptron Neural Network with Selective-Data Training for Flight Arrival Delay Prediction. *Scientific Programming*, *2021*, 1–12. https://doi.org/10.1155/2021/5558918

Cahyadi, Purnomo, D., Nasution, D. S., & Anggraini, F. (2025). Penilaian Esai Mata Kuliah Bahasa Inggris Mahasiswa Berbasis Machine Learning Menggunakan Algoritma Regresi Linier. *Infotech Journal*, *11*(1), 68–72.

Chamidah, N., & Santoni, M. M. (2021). Pencocokan Berbasis Kata Kunci pada Penilaian Esai Pendek Otomatis Berbahasa Indonesia. *Techno.COM*, *20*(1), 19–27.

Chi, T.-Y., Tang, Y.-M., Lu, C.-W., Zhang, Q.-X., & Jang, J.-S. R. (2023). *WC-SBERT: Zero-Shot Text Classification via SBERT with Self-Training for Wikipedia Categories*. http://arxiv.org/abs/2307.15293

Chicho, B. T., & Bibo Sallow, A. (2021). A Comprehensive Survey of Deep Learning Models Based on Keras Framework. *Journal of Soft Computing and Data Mining*, *2*(2). https://doi.org/10.30880/jscdm.2021.02.02.005

Miharja, M. N. D., & Adhkar, S. (2022). Implementasi Chatbot Deteksi Depresi Dini Pada Mahasiswa Dengan PHQ-9 (Patient Health Questionnaire) Menggunakan NLP (Natural Language Processing). *Seminar Nasional Sains Dan Teknologi (SAINTEK)*, *1*(1), 103–108.

Ouahi, M., Khoulji, S., & Laarbi Kerkeb, M. (2024). Analysis of Deep Learning Development Platforms and Their Applications in Sustainable Development within the Education Sector. *E3S Web of Conferences*, *477*(98), 1–11. https://doi.org/10.1051/e3sconf/202447700098

Pangkasidhi, M. K., Palit, H. N., & Gunawan, A. (2022). Analisis Sentimen Mahasiswa di Surabaya Terhadap Pelayanan Vaksinasi COVID-19 Menggunakan Beberapa Classifier. *Jurnal Infra*, *10*(2).

Pradana, R. Y., Nastiti, F. E., & Oktaviani, I. (2024). Machine Learning Pengklasifikasikan Performa Karyawan Direct Sales Force Kartu Prabayar Menggunakan Metode Random Forest Classifier. *JEKIN - Jurnal Teknik Informatika*, *4*(3), 590–599. https://doi.org/10.58794/jekin.v4i3.864

Purnamasari, D., Aji, A. B., Wulandari, D., Reza, F. A., Safrila, M., Yanda, N., & Hidayati, U. (2023). *Pengantar Metode Analisis Sentimen* (1st ed.). Penerbit Gunadarma.

Putra, F. P. E., & Asmunin. (2024). Penerapan Natural Language Processing Pada Aplikasi Chatbot Sebagai Asisten Virtual UMKM Hafidz Pigura. *Jurnal Manajemen Informatika*, *13*(2), 1–9.

Putra, I. F. (2019). *Indonesian Essay Scoring using Bi-LSTM with Word Embedding Representation*.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14

Reimers, N., & Gurevych, I. (2019). 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

Rinjeni, T. P., Indriawan, A., & Rakhmawati, N. A. (2024). Matching Scientific Article Titles using Cosine Similarity and Jaccard Similarity Algorithm. *Procedia Computer Science*, *234*, 553–560. https://doi.org/10.1016/j.procs.2024.03.039

Rumaisa, F., Puspitarani, Y., Rosita, A., Zakiah, A., & Violina, S. (2021). Penerapan Natural Language Processing (NLP) di bidang pendidikan. *Jurnal Inovasi Masyarakat*, *1*(3), 232–235. https://doi.org/10.33197/jim.vol1.iss3.2021.799

Santander-Cruz, Y., Salazar-Colores, S., Paredes-García, W. J., Guendulain-Arenas, H., & Tovar-Arriaga, S. (2022). Semantic Feature Extraction Using SBERT for Dementia Detection. *Brain Sciences*, *12*(2), 270. https://doi.org/10.3390/brainsci12020270

Setio Pribadi, F., Hasanah, U., & Nur Sa, D. (2023, July 13). Automatic Short Answer Scoring (ASAS) Using String-based Similarity and Query Expansion. *Proceedings of Vocational Engineering International Conference*.

Yang, R., Cao, J., Wen, Z., Wu, Y., & He, X. (2020). Enhancing Automated Essay Scoring Performance via Fine-tuning Pre-trained Language Models with Combination of Regression and Ranking. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1560–1569. https://doi.org/10.18653/v1/2020.findings-emnlp.141

Zupon, A., Carnie, A., Hammond, M., & Surdeanu, M. (2022). *Automatic Correction of Syntactic Dependency Annotation Differences*. http://arxiv.org/abs/2201.05891