Automated Essay Scoring menggunakan *Sentence Embedding*, *Linguistic Features* dan *Random Forest Regression*

Hilmiyatul Asna
Teknik Informatika, Teknik Informatika dan Komputer, Politeknik Negeri Jakarta
Depok, Jawa Barat
hilmiyatul.asna.tik22@mhsw.pnj.ac.id

Abstract* - Essay assessment plays an essential role in evaluating students' critical thinking and writing abilities. However, manual essay grading is time-consuming, labor-intensive, and susceptible to subjective judgment, particularly in classrooms with a large number of students. Recent Automated Essay Scoring (AES) studies have demonstrated promising results using semantic text representations or handcrafted linguistic features independently. Nevertheless, limited research has explored the integration of sentence embedding and linguistic features within a machine learning framework for Indonesian multilingual essay assessment while simultaneously implementing the model in a practical web-based system. This study proposes an Automated Essay Scoring system that combines semantic representations generated by the paraphrase-multilingual-MiniLM-L12-v2 sentence embedding model with fourteen linguistic features and Random Forest Regression for essay score prediction. The proposed system was developed as a web-based application using Flask and evaluated using a dataset consisting of 750 manually scored student essays from three subjects: Indonesian Language, Civic Education, and Islamic History. The extracted features include sentence embeddings, cosine similarity, linguistic features, named entity recognition, word sense disambiguation, language detection, and one-hot encoded question identifiers. Experimental results demonstrate that the proposed model achieved a Quadratic Weighted Kappa (QWK) of 0.8552, RMSE of 13.71, MAE of 8.64, R^2 of 0.7781, Pearson correlation of 0.9018, and Spearman correlation of 0.8443, indicating a high level of agreement between automated and manual scoring. Feature importance analysis further reveals that sentence embedding contributes 78.04% of the predictive capability, while linguistic features contribute 19.09%, confirming the effectiveness of combining semantic and linguistic information. The proposed system provides an efficient, objective, and practical solution for supporting essay assessment in educational environments.

Keywords: Automated Essay Scoring, Sentence Embedding, Linguistic Features, Random Forest Regression, Natural Language Processing.

Abstrak*-- Penilaian esai merupakan salah satu komponen penting dalam mengevaluasi kemampuan berpikir kritis dan keterampilan menulis siswa. Namun, proses penilaian secara manual membutuhkan waktu yang lama, memerlukan tenaga yang besar, serta rentan terhadap subjektivitas, terutama ketika jumlah esai yang dinilai cukup banyak. Berbagai penelitian mengenai Automated Essay Scoring (AES) telah menunjukkan bahwa representasi semantik maupun fitur linguistik mampu meningkatkan kualitas prediksi skor esai. Akan tetapi, penelitian yang mengintegrasikan sentence embedding dan linguistic features dalam kerangka machine learning untuk penilaian esai berbahasa Indonesia serta mengimplementasikannya pada sistem berbasis web masih terbatas. Penelitian ini mengusulkan sistem Automated Essay Scoring yang mengombinasikan representasi semantik menggunakan model paraphrase-multilingual-MiniLM-L12-v2, empat belas fitur linguistik, dan algoritma Random Forest Regression untuk memprediksi skor esai siswa. Sistem dikembangkan sebagai aplikasi berbasis web menggunakan Flask dan dievaluasi menggunakan 750 jawaban esai siswa yang telah dinilai guru pada tiga mata pelajaran, yaitu Bahasa Indonesia, Pendidikan Kewarganegaraan, dan Sejarah Islam. Fitur yang digunakan meliputi sentence embedding, cosine similarity, fitur linguistik, named entity recognition, word sense disambiguation, deteksi bahasa, serta one-hot encoding identitas soal. Hasil eksperimen menunjukkan bahwa model memperoleh nilai Quadratic Weighted Kappa (QWK) sebesar 0,8552, RMSE sebesar 13,71, MAE sebesar 8,64, R^2 sebesar 0,7781, Pearson Correlation sebesar 0,9018, dan Spearman Correlation sebesar 0,8443, yang menunjukkan tingkat kesesuaian yang tinggi terhadap penilaian guru. Analisis feature importance menunjukkan bahwa sentence embedding memberikan kontribusi terbesar sebesar 78,04%, sedangkan linguistic features berkontribusi sebesar 19,09%, sehingga kombinasi keduanya mampu meningkatkan kualitas prediksi. Hasil penelitian menunjukkan bahwa sistem yang diusulkan mampu

memberikan proses penilaian esai yang lebih efisien, konsisten, dan objektif untuk mendukung evaluasi pembelajaran berbasis teknologi.

Kata kunci: Automated Essay Scoring, Sentence Embedding, Linguistic Features, Random Forest Regression, Natural Language Processing.

I. PENDAHULUAN

Penilaian esai merupakan salah satu bentuk evaluasi pembelajaran yang mampu mengukur kemampuan berpikir kritis, pemahaman konsep, kemampuan bernalar, serta keterampilan menulis siswa secara lebih komprehensif dibandingkan soal objektif. Melalui jawaban esai, siswa dituntut untuk mengembangkan ide, menyusun argumentasi, dan menghubungkan konsep menggunakan bahasa mereka sendiri sehingga guru dapat memperoleh gambaran yang lebih utuh mengenai tingkat penguasaan materi. Oleh karena itu, penilaian esai masih menjadi salah satu metode evaluasi yang banyak diterapkan pada berbagai mata pelajaran, khususnya yang menekankan kemampuan analisis dan komunikasi tertulis [1], [2].

Meskipun memiliki kelebihan dalam mengukur kemampuan kognitif tingkat tinggi, proses penilaian esai secara manual masih menghadapi berbagai kendala. Guru memerlukan waktu yang cukup lama untuk membaca dan mengevaluasi setiap jawaban siswa, terutama ketika jumlah peserta didik yang dinilai relatif banyak. Selain itu, proses penilaian sering dipengaruhi oleh subjektivitas, kelelahan penilai, maupun perbedaan interpretasi terhadap rubrik penilaian sehingga dapat menghasilkan skor yang kurang konsisten. Kondisi tersebut menyebabkan proses evaluasi menjadi kurang efisien dan meningkatkan beban kerja guru [3], [4].

Perkembangan Artificial Intelligence (AI) dan Natural Language Processing (NLP) telah mendorong lahirnya berbagai penelitian mengenai *Automated Essay Scoring* (AES), yaitu sistem yang mampu memberikan skor terhadap jawaban esai secara otomatis berdasarkan model yang dipelajari dari data yang telah diberi penilaian oleh guru. Penggunaan AES bertujuan untuk meningkatkan efisiensi proses penilaian sekaligus menjaga konsistensi hasil evaluasi. Berbagai pendekatan telah dikembangkan, mulai dari metode berbasis frekuensi kata seperti *Term Frequency–Inverse Document Frequency* (TF-IDF), representasi berbasis *word embedding*, hingga model transformer yang mampu memahami hubungan semantik antar kalimat secara lebih kontekstual [5], [6].

Salah satu pendekatan yang banyak digunakan pada penelitian AES modern adalah Sentence Embedding. Teknik ini merepresentasikan setiap kalimat ke dalam bentuk vektor numerik yang mempertahankan informasi semantik sehingga sistem mampu mengenali jawaban yang memiliki makna serupa meskipun menggunakan pilihan kata atau struktur kalimat yang berbeda. Dibandingkan metode representasi teks berbasis frekuensi kata, *Sentence Embedding* memberikan representasi yang lebih kaya karena mempertimbangkan konteks kalimat secara keseluruhan. Oleh karena itu, pendekatan ini dinilai lebih sesuai untuk tugas penilaian esai yang menekankan kesesuaian makna dibandingkan kemunculan kata secara eksplisit [7], [8].

Selain kesamaan makna, kualitas jawaban esai juga dipengaruhi oleh karakteristik kebahasaan yang digunakan siswa. Dalam praktik penilaian manual, guru tidak hanya mempertimbangkan kesesuaian isi jawaban terhadap kunci jawaban, tetapi juga memperhatikan struktur kalimat, variasi kosakata, panjang jawaban, serta penggunaan unsur kebahasaan lainnya. Informasi tersebut dapat direpresentasikan melalui Linguistic Features sehingga mampu memberikan informasi tambahan mengenai kualitas penulisan yang belum sepenuhnya direpresentasikan oleh *Sentence Embedding*. Dengan mengkombinasikan kedua jenis fitur tersebut, model diharapkan mampu menangkap aspek semantik sekaligus kualitas kebahasaan secara bersamaan [9], [10].

A. Penelitian Terdahulu

Sebelum penelitian ini dilakukan, berbagai pendekatan telah dikembangkan untuk meningkatkan performa *Automated Essay Scoring* (AES). Penelitian-penelitian tersebut memanfaatkan berbagai algoritma *machine learning* maupun *deep learning*, seperti transformer, *feature engineering*, dan *sentence embedding*, untuk meningkatkan kemampuan model dalam memahami jawaban esai siswa. Meskipun demikian, masih terdapat keterbatasan dalam mengintegrasikan representasi semantik dan karakteristik linguistik secara bersamaan, khususnya pada penilaian esai berbahasa Indonesia. Ringkasan beberapa penelitian terdahulu yang relevan disajikan pada Tabel I.

TABEL I. PERBANDINGAN PENELITIAN TERDAHULU

Peneliti	Metode	Keterbatasan
Rahmat et al. (2025)	DeBERTa-v3	Belum mengombinasikan fitur linguistik sebagai representasi tambahan.
Rafrin et al. (2025)	Feature Engineering + SVM	Belum menggunakan representasi semantik berbasis *Sentence Embedding*.
Sitorus et al. (2026)	Sentence-BERT + Hybrid Similarity	Tidak menggunakan model regresi dan belum mengintegrasikan fitur linguistik.
Penelitian ini	Sentence Embedding + Linguistic Features + Random Forest Regression	-

B. Research Gap

Berdasarkan penelitian terdahulu pada Tabel I, masih terdapat beberapa kesenjangan penelitian yang dapat dikembangkan. Pertama, sebagian besar penelitian hanya berfokus pada penggunaan representasi semantik berbasis transformer atau fitur linguistik secara terpisah sehingga informasi semantik dan karakteristik kebahasaan belum dimanfaatkan secara optimal dalam proses prediksi skor. Kedua, penelitian yang menerapkan Random Forest Regression sebagai model prediksi dengan mengombinasikan *Sentence Embedding* dan *Linguistic Features* pada penilaian esai berbahasa Indonesia masih sangat terbatas. Ketiga, sebagian besar penelitian hanya mengevaluasi performa model tanpa mengintegrasikannya ke dalam sistem berbasis web yang dapat digunakan secara langsung oleh guru dalam proses penilaian. Oleh karena itu, penelitian ini mengusulkan integrasi *Sentence Embedding*, *Linguistic Features*, dan *Random Forest Regression* ke dalam sistem *Automated Essay Scoring* berbasis web untuk menghasilkan proses penilaian yang lebih akurat, konsisten, dan mudah diimplementasikan dalam lingkungan pendidikan.

C. Kontribusi Penelitian

Berdasarkan *research gap* tersebut, penelitian ini mengusulkan sistem Automated Essay Scoring berbasis web yang mengkombinasikan *Sentence Embedding* menggunakan model *paraphrase-multilingual-MiniLM-L12-v2* dengan empat belas Linguistic Features menggunakan algoritma Random Forest Regression. Kontribusi utama penelitian ini meliputi:

1. Mengintegrasikan *Sentence Embedding* dan empat belas Linguistic Features sebagai representasi semantik dan karakteristik kebahasaan dalam proses penilaian esai otomatis.
2. Menerapkan algoritma Random Forest Regression untuk memprediksi skor esai berdasarkan kombinasi kedua kelompok fitur tersebut.
3. Mengembangkan sistem Automated Essay Scoring berbasis web menggunakan framework Flask sehingga model dapat dimanfaatkan secara langsung dalam proses pembelajaran oleh guru dan siswa.
4. Mengevaluasi performa model menggunakan Quadratic Weighted Kappa (QWK), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2, Pearson Correlation, dan Spearman Correlation, serta menganalisis kontribusi setiap kelompok fitur melalui analisis *feature importance*.

Hasil penelitian ini diharapkan dapat memberikan alternatif solusi penilaian esai yang lebih cepat, objektif, dan konsisten sekaligus memberikan kontribusi terhadap pengembangan penerapan Artificial Intelligence dan Natural Language Processing pada bidang teknologi pendidikan.

II. METODOLOGI PENELITIAN

Penelitian ini mengusulkan sistem *Automated Essay Scoring* (AES) berbasis web yang mengkombinasikan representasi semantik menggunakan *Sentence Embedding* dengan karakteristik kebahasaan melalui *Linguistic Features*. Seluruh fitur yang dihasilkan digunakan sebagai masukan bagi algoritma *Random Forest Regression* untuk memprediksi skor jawaban esai siswa. Tahapan metode yang diusulkan meliputi penyusunan dataset, *preprocessing*, ekstraksi fitur, pelatihan model, serta evaluasi performa model.

A. Dataset

Dataset penelitian terdiri atas 750 jawaban esai siswa yang diperoleh dari MTs Amal Bakti. Data berasal dari tiga mata pelajaran, yaitu Bahasa Indonesia, Pendidikan Kewarganegaraan (PKN),

dan Sejarah Islam. Seluruh jawaban telah diberi skor oleh guru sehingga dapat digunakan sebagai *ground truth* dalam proses pelatihan dan evaluasi model.

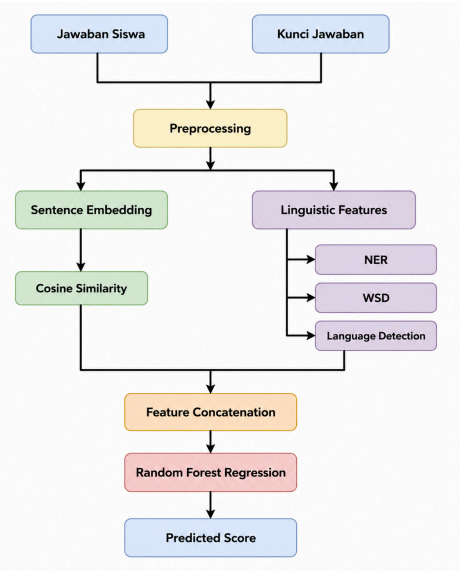
Setiap data terdiri atas identitas soal (*soal_id*), mata pelajaran, jawaban siswa, kunci jawaban, dan skor penilaian guru. Dataset dibagi menjadi 675 data latih dan 75 data uji. Ringkasan dataset ditunjukkan pada Tabel II.

TABEL II. RINGKASAN DATASET PENELITIAN

Parameter	Nilai
Jumlah dataset	750
Data latih	675
Data uji	75
Mata pelajaran	3
Ground truth	Skor guru

B. Metode yang Diusulkan

Arsitektur metode yang diusulkan ditunjukkan pada Gambar 1. Sistem menerima dua masukan berupa jawaban siswa dan kunci jawaban. Kedua teks terlebih dahulu melalui tahap *preprocessing* untuk menghasilkan representasi yang konsisten. Selanjutnya dilakukan ekstraksi fitur melalui dua jalur utama, yaitu representasi semantik menggunakan *Sentence Embedding* dan representasi kebahasaan menggunakan *Linguistic Features*. Seluruh fitur kemudian digabungkan melalui proses *feature concatenation* sebelum diproses oleh model *Random Forest Regression* untuk menghasilkan skor prediksi.



Gambar 1. Arsitektur Metode Automated Essay Scoring yang Diusulkan

TABEL III. REPRESENTASI FITUR YANG DIGUNAKAN

Kelompok Fitur	Jumlah
Sentence Embedding (Jawaban Siswa)	384
Sentence Embedding (Kunci Jawaban)	384
Cosine Similarity	1
Linguistic Features	14
NER	5
WSD	3
Language Detection	1
One-Hot Encoding (*soal_id*)	5
Total Fitur	798

Tahapan pada metode yang diusulkan dijelaskan sebagai berikut.

1) Preprocessing

Tahap *preprocessing* bertujuan menghasilkan format teks yang seragam sebelum dilakukan ekstraksi fitur. Proses yang dilakukan meliputi *lowercasing*, penghapusan karakter yang tidak diperlukan, serta normalisasi spasi. Berbeda dengan pendekatan klasifikasi teks konvensional, penelitian ini tidak menerapkan *stemming* maupun *stopword removal* agar informasi semantik tetap dipertahankan selama proses pembentukan *sentence embedding*.

2) Sentence Embedding

Representasi semantik diperoleh menggunakan model paraphrase-multilingual-MiniLM-L12-v2 dari pustaka *Sentence Transformers*. Model ini menghasilkan representasi vektor berdimensi 384 untuk setiap kalimat dan mendukung berbagai bahasa, termasuk Bahasa Indonesia. Pada penelitian ini, *embedding* dibentuk untuk jawaban siswa dan kunci jawaban sehingga diperoleh total 768 fitur embedding.

Selain vektor *embedding*, dihitung pula nilai Cosine Similarity sebagai ukuran tingkat kesamaan makna antara jawaban siswa dan kunci jawaban.

3) Linguistic Features

Selain representasi semantik, penelitian ini mengekstraksi 14 fitur linguistik yang menggambarkan karakteristik kebahasaan jawaban

siswa. Fitur tersebut meliputi jumlah karakter, jumlah kata, jumlah kalimat, rata-rata panjang kalimat, variasi kosakata, rasio kelas kata, kepadatan leksikal, struktur SPOK, serta indikator keterbacaan.

Untuk memperkaya representasi teks, penelitian ini juga mengintegrasikan Named Entity Recognition (NER), Word Sense Disambiguation (WSD), Language Detection, serta One-Hot Encoding pada atribut *soal_id*. Seluruh fitur kemudian digabungkan menjadi satu vektor fitur menggunakan proses *feature concatenation* sebelum digunakan sebagai masukan model prediksi.

4) Random Forest Regression

Prediksi skor esai dilakukan menggunakan algoritma Random Forest Regression, yaitu metode *ensemble learning* yang membangun sejumlah *decision tree* dan menghasilkan prediksi berdasarkan rata-rata keluaran seluruh pohon keputusan. Algoritma ini dipilih karena mampu menangani data berdimensi tinggi, memiliki ketahanan terhadap *overfitting*, serta menyediakan analisis *feature importance* yang dapat digunakan untuk mengetahui kontribusi masing-masing fitur terhadap hasil prediksi.

Parameter terbaik yang digunakan pada penelitian ini ditunjukkan pada Tabel IV.

TABEL IV. PARAMETER RANDOM FOREST REGRESSION

Parameter	Nilai
n_estimators	700
max_features	sqrt
max_depth	None
random_state	42

C. Pengaturan Eksperimen dan Evaluasi

Model dilatih menggunakan data latih dan selanjutnya diuji menggunakan 75 data uji yang tidak terlibat selama proses pelatihan. Performa model dievaluasi menggunakan Quadratic Weighted Kappa (QWK) untuk mengukur tingkat kesesuaian antara skor sistem dan skor guru, Root Mean Squared Error (RMSE) dan Mean Absolute Error (MAE) untuk mengukur kesalahan prediksi, koefisien determinasi (R^2) untuk mengukur kemampuan model menjelaskan variasi data, serta Pearson Correlation dan Spearman Correlation untuk mengevaluasi hubungan antara hasil prediksi sistem dengan penilaian guru.

Selain evaluasi model, aplikasi yang dikembangkan juga diuji melalui Black Box Testing, System Usability Scale (SUS), Net Promoter Score (NPS), dan User Acceptance Test (UAT) untuk memastikan bahwa sistem memenuhi aspek fungsionalitas, kemudahan penggunaan, serta penerimaan pengguna.

III. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil eksperimen terhadap model *Automated Essay Scoring* yang diusulkan serta pembahasannya. Evaluasi dilakukan untuk mengetahui kemampuan model dalam memprediksi skor jawaban esai siswa berdasarkan kombinasi *Sentence Embedding* dan *Linguistic Features* menggunakan algoritma *Random Forest Regression*. Selain mengevaluasi performa model, penelitian ini juga menganalisis kontribusi masing-masing kelompok fitur terhadap hasil prediksi serta implementasi model pada sistem berbasis web.

A. Hasil Evaluasi Model

Evaluasi model dilakukan menggunakan 75 data uji yang tidak digunakan selama proses pelatihan. Kinerja model diukur menggunakan Quadratic Weighted Kappa (QWK), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), koefisien determinasi (R^2), Pearson Correlation, dan Spearman Correlation. Hasil evaluasi ditunjukkan pada Tabel V.

TABEL V. HASIL EVALUASI MODEL

Metrik	Nilai
Quadratic Weighted Kappa	0.8552
RMSE	13.7060
MAE	8.6405
R^2	0.7781
Pearson Correlation	0.9018
Spearman Correlation	0.8443

Nilai Quadratic Weighted Kappa (QWK) sebesar 0.8552 menunjukkan tingkat kesesuaian yang tinggi antara skor yang dihasilkan sistem dan skor yang diberikan guru. Berdasarkan interpretasi Landis dan Koch, nilai tersebut termasuk kategori *Almost Perfect Agreement*, yang menunjukkan bahwa model mampu merepresentasikan pola penilaian guru secara konsisten.

Nilai RMSE sebesar 13.7060 dan MAE sebesar 8.6405 menunjukkan bahwa rata-rata

kesalahan prediksi relatif kecil terhadap rentang skor penilaian yang digunakan. Selain itu, nilai R^2 sebesar 0.7781 menunjukkan bahwa sekitar 77,81% variasi skor penilaian guru dapat dijelaskan oleh model yang diusulkan.

Hubungan antara skor prediksi sistem dan skor manual guru juga menunjukkan korelasi yang sangat kuat. Hal tersebut ditunjukkan oleh nilai Pearson Correlation sebesar 0.9018 dan Spearman Correlation sebesar 0.8443, yang mengindikasikan bahwa model tidak hanya mampu menghasilkan prediksi yang mendekati nilai guru secara numerik, tetapi juga mempertahankan urutan peringkat skor secara konsisten.

B. Analisis Kesesuaian Prediksi

Selain menggunakan metrik evaluasi, penelitian ini juga menganalisis tingkat kesesuaian antara skor prediksi sistem dan skor manual guru berdasarkan selisih skor. Distribusi selisih prediksi ditunjukkan pada Tabel VI.

TABEL VI. DISTRIBUSI SELISIH PREDIKSI

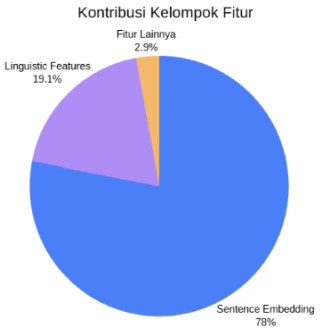
Selisih	Jumlah	Persentase
0	5	6.67%
1-5	33	44.00%
6-10	19	25.33%
>10	18	24.00%

Sebanyak 50,67% prediksi sistem memiliki selisih maksimal lima poin terhadap skor guru, sedangkan 72% prediksi memiliki selisih tidak lebih dari sepuluh poin. Hasil tersebut menunjukkan bahwa sebagian besar prediksi sistem berada pada rentang penilaian yang masih mendekati hasil penilaian manual guru.

Meskipun demikian, masih terdapat 24% data yang memiliki selisih lebih dari sepuluh poin. Berdasarkan hasil observasi terhadap data tersebut, sebagian besar kesalahan terjadi pada jawaban siswa yang memiliki struktur kalimat tidak lengkap, penggunaan istilah yang ambigu, atau jawaban yang sangat singkat sehingga informasi semantik yang diperoleh model menjadi terbatas.

C. Analisis Kontribusi Fitur

Untuk mengetahui kontribusi masing-masing kelompok fitur terhadap hasil prediksi, dilakukan analisis *feature importance* pada model *Random Forest Regression*. Hasil analisis ditunjukkan pada Gambar 3.



Gambar 3. Kontribusi Kelompok Fitur terhadap Model

Hasil analisis menunjukkan bahwa Sentence Embedding memberikan kontribusi terbesar terhadap proses prediksi dengan persentase 78,04%. Hal ini menunjukkan bahwa representasi semantik memiliki peran dominan dalam menentukan skor jawaban esai karena mampu menangkap hubungan makna antara jawaban siswa dan kunci jawaban.

Di sisi lain, Linguistic Features memberikan kontribusi sebesar 19,09%. Meskipun kontribusinya lebih kecil dibandingkan *Sentence Embedding*, fitur linguistik tetap memberikan informasi penting mengenai kualitas kebahasaan, seperti struktur kalimat, variasi kosakata, serta karakteristik penulisan siswa yang tidak sepenuhnya direpresentasikan oleh embedding.

Kontribusi sisanya berasal dari fitur pendukung seperti Cosine Similarity, Named Entity Recognition (NER), Word Sense Disambiguation (WSD), dan Language Detection, yang membantu memperkaya representasi data meskipun pengaruhnya relatif kecil terhadap hasil prediksi akhir.

D. Perbandingan dengan Penelitian Terdahulu

Untuk mengetahui efektivitas metode yang diusulkan, hasil penelitian dibandingkan dengan beberapa penelitian sebelumnya yang menerapkan pendekatan *Automated Essay Scoring*. Perbandingan tersebut disajikan pada Tabel VII.

TABEL VII. PERBANDINGAN DENGAN PENELITIAN TERDAHULU

Penelitian	Metode	Metrik Evaluasi
Rahmat et al. (2025)	DeBERTa-v3	QWK = 0.558
Rafrin et al. (2025)	Feature Engineering + SVM	QWK = 0.8397

Penelitian	Metode	Metrik Evaluasi
Sitorus et al. (2026)	Sentence-BERT + Hybrid Similarity	Agreement 87.3%
Penelitian ini	Sentence Embedding + Linguistic Features + Random Forest Regression	QWK = 0.8552

Berdasarkan hasil tersebut, metode yang diusulkan menunjukkan performa yang kompetitif dibandingkan beberapa penelitian sebelumnya. Integrasi antara *Sentence Embedding* dan *Linguistic Features* mampu meningkatkan kemampuan model dalam memahami makna jawaban sekaligus mempertimbangkan kualitas kebahasaan sehingga menghasilkan tingkat kesesuaian yang tinggi terhadap penilaian guru.

E. Implementasi Sistem

Model yang telah dilatih kemudian diintegrasikan ke dalam aplikasi *Automated Essay Scoring* berbasis web menggunakan framework Flask. Sistem memungkinkan guru membuat mata pelajaran, mengelola ujian esai, serta memperoleh hasil penilaian secara otomatis setelah siswa mengirimkan jawaban. Selain menghasilkan skor prediksi, sistem juga menyimpan hasil penilaian ke dalam basis data sehingga dapat digunakan sebagai bahan evaluasi pembelajaran.

Pengujian fungsional menunjukkan bahwa seluruh fitur sistem berjalan sesuai dengan kebutuhan pengguna. Selain itu, hasil pengujian System Usability Scale (SUS) memperoleh skor 82,5, yang termasuk kategori Excellent, sedangkan hasil Net Promoter Score (NPS) menunjukkan nilai 60%, yang mengindikasikan tingkat kepuasan pengguna yang baik. Hasil User Acceptance Test (UAT) juga menunjukkan tingkat penerimaan sebesar 97,6% oleh guru dan 94,4% oleh siswa, sehingga sistem dinilai layak digunakan dalam mendukung proses penilaian esai di lingkungan sekolah.

IV. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengembangkan sistem *Automated Essay Scoring* (AES) berbasis web dengan mengintegrasikan *Sentence Embedding*, *Linguistic Features*, dan algoritma *Random Forest Regression* untuk memprediksi skor jawaban esai siswa. Kombinasi representasi semantik menggunakan model *paraphrase-multilingual-MiniLM-L12-v2* dengan empat belas fitur linguistik mampu menghasilkan performa prediksi yang baik, ditunjukkan oleh nilai Quadratic Weighted Kappa (QWK) sebesar 0,8552, RMSE sebesar 13,7060, MAE sebesar 8,6405, R^2 sebesar 0,7781, Pearson Correlation sebesar 0,9018, dan Spearman Correlation sebesar 0,8443. Hasil tersebut menunjukkan bahwa model memiliki tingkat kesesuaian yang tinggi terhadap penilaian manual guru. Selain itu, analisis *feature importance* menunjukkan bahwa *Sentence Embedding* memberikan kontribusi terbesar terhadap proses prediksi, sedangkan *Linguistic Features* berperan sebagai fitur pelengkap yang meningkatkan kemampuan model dalam merepresentasikan kualitas kebahasaan jawaban siswa. Integrasi model ke dalam aplikasi berbasis web juga menunjukkan bahwa sistem dapat digunakan sebagai alternatif yang lebih cepat, konsisten, dan objektif dalam mendukung proses penilaian esai di lingkungan pendidikan.

Meskipun demikian, penelitian ini masih memiliki beberapa keterbatasan. Dataset yang digunakan berasal dari satu institusi pendidikan dengan tiga mata pelajaran sehingga kemampuan generalisasi model terhadap mata pelajaran atau jenjang pendidikan lain masih perlu dikaji lebih lanjut. Selain itu, penelitian ini belum mengeksplorasi pendekatan berbasis *Large Language Models* (LLMs) atau proses *fine-tuning* model transformer yang berpotensi meningkatkan kemampuan pemahaman konteks jawaban siswa. Oleh karena itu, penelitian selanjutnya dapat memperluas cakupan dataset dari berbagai sekolah, menguji performa pada mata pelajaran lain, membandingkan metode yang diusulkan dengan model transformer terkini, serta mengembangkan mekanisme penilaian yang tidak hanya menghasilkan skor akhir tetapi juga memberikan umpan balik (*feedback*) otomatis kepada siswa mengenai kualitas jawaban yang diberikan.

V. REFERENSI

[1] Z. Ke and V. Ng, “Automated Essay Scoring: A Survey of the State of the Art,” *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, Macau, China, 2019, pp. 6300–6308.

[2] M. A. Hussein, H. Hassan, and M. Nassef, “Automated Language Essay Scoring Systems: A Literature Review,” *PeerJ Computer Science*, vol. 5, p. e208, 2019.

[3] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 3982–3992.

[4] F. Wang, W. Yang, M. Neumann, et al., “paraphrase-multilingual-MiniLM-L12-v2,” *Sentence-Transformers*, Hugging Face, 2021.

[5] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., Pearson, 2024.

[7] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[8] M. M. Mukaka, “Statistics Corner: A Guide to Appropriate Use of Correlation Coefficient in Medical Research,” *Malawi Medical Journal*, vol. 24, no. 3, pp. 69–71, 2012.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[10] M. Grinberg, *Flask Web Development: Developing Web Applications with Python*, 2nd ed., Sebastopol: O’Reilly Media, 2018.

[11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of NAACL-HLT*, Minneapolis, USA, 2019, pp. 4171–4186.

[12] F. Nadeem, H. Nguyen, Y. Liu, and M. Ostendorf, “Automated Essay Scoring with Discourse-Aware Neural Models,” *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Florence, Italy, 2019, pp. 484–493.

[13] R. Navigli, “Word Sense Disambiguation: A Survey,” *ACM Computing Surveys*, vol. 41, no. 2, pp. 1–69, 2009.

[14] Z. Ke and V. Ng, “Recent Advances in Automated Essay Scoring: A Review of Transformer-Based Methods,” *IEEE Access*, 2024.

[15] Q. Faseeh, et al., “Automated Essay Scoring Using RoBERTa Embeddings and Handcrafted Linguistic Features with Lightweight XGBoost,” *Mathematics*, vol. 12, no. 21, 2024.

[16] H. Firoozi, et al., “Using Automated Procedures to Score Educational Essays Written in Three Languages,” *Journal of Educational Measurement*, vol. 62, no. 1, 2025.

[17] A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention Is All You Need,” *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, USA, 2017.

[18] Y. Liu, M. Ott, N. Goyal, et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv:1907.11692, 2019.

[19] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation,” *Proceedings of EMNLP*, Doha, Qatar, 2014, pp. 1532–1543.

[20] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and Their Compositionality,” *Advances in Neural Information Processing Systems (NeurIPS)*, Lake Tahoe, USA, 2013.