JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 218-227

Analisis Komparatif Metode Bag Of Words, TF-IDF, dan Transformer pada Sistem Penilaian Esai Otomatis Berbasis Kecerdasan Buatan

Amara Seviany Agustin[1], Suprih Widodo[2], Ulva Elviani [3], Ayu Permata Sari [4], Muhamad Akda Fathul Barri [5]

[12345]Program Studi Pendidikan Sistem dan Teknologi Informasi, Universitas Pendidikan Indonesia, Bandung, Indonesia
Email: [1]amara.seviany8@upi.edu*, [2]supri@upi.edu, [3]ulva@upi.edu, [4]ayupermata29@upi.edu, [5]akdafathul@upi.edu
(* : coressponding author)

Abstrak−Penilaian esai secara manual menghadapi kendala inkonsistensi, subjektivitas, dan keterbatasan waktu, terutama pada pembelajaran berskala besar. Penelitian ini membandingkan tiga pendekatan representasi teks pada sistem penilaian esai otomatis berbasis kecerdasan buatan, yaitu Bag of Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), dan Transformer (IndoBERT). Dataset yang digunakan berasal dari Kaggle Learning Agency Lab Automated Essay Scoring 2.0 yang terdiri atas 17.207 esai berbahasa Inggris dan diterjemahkan ke bahasa Indonesia menggunakan model Helsinki-NLP opus-mt-en-id. Tahap prapemrosesan meliputi case folding, pembersihan teks, penghapusan stopword, dan stemming menggunakan pustaka Sastrawi. Metode BoW dan TF-IDF dipadukan dengan Support Vector Regression, sedangkan pendekatan Transformer menggunakan fine-tuning IndoBERT. Evaluasi dilakukan menggunakan metrik Quadratic Weighted Kappa (QWK). Hasil eksperimen menunjukkan bahwa IndoBERT mencapai performa tertinggi dengan nilai QWK sebesar 0,7842, diikuti TF-IDF sebesar 0,6521 dan BoW sebesar 0,6103. Meskipun Transformer unggul dari sisi akurasi, metode klasik tetap relevan untuk implementasi dengan keterbatasan sumber daya komputasi karena efisiensi waktu dan kompleksitas yang lebih rendah. Temuan ini menegaskan pentingnya pemilihan metode penilaian otomatis yang disesuaikan dengan konteks kebutuhan dan infrastruktur pendidikan.

Kata Kunci: automated essay scoring; bag of words; IndoBERT; TF-IDF; transformer

Abstract−Manual essay assessment faces challenges of inconsistency, subjectivity, and time constraints, particularly in large-scale learning environments. This study compares three text representation approaches for automated essay scoring systems: Bag of Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), and Transformer (IndoBERT). The dataset is obtained from the Kaggle Learning Agency Lab Automated Essay Scoring 2.0, consisting of 17,207 English essays translated into Indonesian using the Helsinki-NLP opus-mt-en-id model. Preprocessing stages include case folding, text cleaning, stopword removal, and stemming using the Sastrawi library. Classical approaches (BoW and TF-IDF) are combined with Support Vector Regression, while the Transformer-based approach applies fine-tuned IndoBERT. Evaluation using Quadratic Weighted Kappa (QWK) indicates that IndoBERT achieves the highest performance with a QWK of 0.7842, followed by TF-IDF at 0.6521 and BoW at 0.6103. Although Transformer models demonstrate superior accuracy, classical methods remain practical for resource-constrained implementations due to their lower computational complexity and faster inference time

Keywords: automated essay scoring; bag of words; IndoBERT; TF-IDF; transformer

1. PENDAHULUAN

Perkembangan kecerdasan buatan telah membawa perubahan signifikan dalam berbagai aspek pendidikan, khususnya pada proses evaluasi pembelajaran (Russell & Bennett, 2020). Penilaian tugas esai merupakan salah satu komponen penting dalam evaluasi hasil belajar karena mampu mengukur kemampuan berpikir kritis, analitis, dan argumentatif peserta didik (Kumar & Bakhshi, 2018). Namun demikian, penilaian esai secara manual memerlukan waktu yang panjang, biaya yang tinggi, serta berpotensi menimbulkan subjektivitas dan inkonsistensi antar penilai, terutama pada pembelajaran berskala besar dan berbasis sistem daring (Hussein et al., 2019). Sejumlah studi melaporkan bahwa variabilitas antar penilai tetap muncul bahkan pada penilai yang telah berpengalaman, terutama ketika beban penilaian meningkat (Lane & Stone, 2019). Penelitian oleh Kumar dan Bakhshi (2018) menunjukkan bahwa tingkat kesepakatan antar penilai (inter-rater

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 218-227

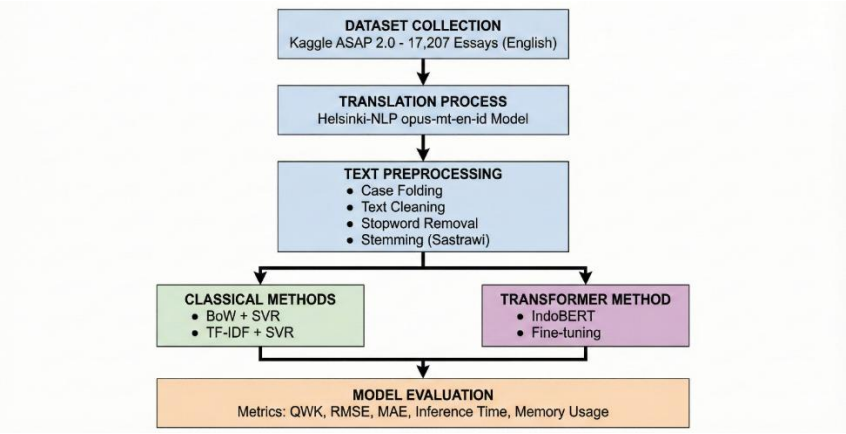
reliability) pada penilaian esai manual berkisar antara 0,60 hingga 0,75 menggunakan metrik Cohen's Kappa, yang mengindikasikan adanya ruang untuk perbaikan konsistensi.

Kondisi ini mendorong pengembangan sistem penilaian esai otomatis (automated essay scoring/AES) sebagai alternatif atau pendukung penilaian konvensional. AES bertujuan untuk meningkatkan efisiensi, konsistensi, dan skalabilitas penilaian tanpa mengabaikan kualitas evaluasi akademik (Shermis & Burstein, 2013). Pendekatan AES telah berkembang dari metode statistik klasik hingga model pembelajaran mendalam berbasis Transformer (Devlin et al., 2019). Metode klasik seperti Bag of Words dan TF-IDF menawarkan kemudahan implementasi, transparansi model, serta efisiensi komputasi (Manning et al., 2008). Penelitian Hussein et al. (2019) menunjukkan bahwa metode TF-IDF dengan Support Vector Machine (SVM) dapat mencapai akurasi hingga 85% pada dataset ASAP dengan waktu pelatihan yang relatif singkat. Sebaliknya, model Transformer seperti BERT dan variannya menunjukkan performa unggul dalam menangkap konteks semantik dan struktur wacana yang kompleks, tetapi menuntut sumber daya komputasi yang jauh lebih besar (Tay et al., 2020).

Dalam konteks pendidikan di Indonesia, keterbatasan infrastruktur komputasi masih menjadi tantangan utama dalam penerapan model deep learning berskala besar (Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, 2023). Data dari Kementerian Pendidikan menunjukkan bahwa hanya 35% institusi pendidikan tinggi di Indonesia memiliki akses ke infrastruktur GPU untuk keperluan komputasi pembelajaran mesin. Oleh karena itu, kajian komparatif yang mengevaluasi trade-off antara akurasi dan efisiensi komputasi menjadi relevan untuk memberikan dasar pengambilan keputusan yang rasional (Zhou & Hovy, 2021). Penelitian ini bertujuan membandingkan performa metode Bag of Words, TF-IDF, dan Transformer (IndoBERT) dalam sistem penilaian esai otomatis berbahasa Indonesia, serta menganalisis implikasi praktis dari masing-masing pendekatan. Kontribusi utama penelitian ini meliputi: (1) evaluasi komprehensif tiga paradigma representasi teks pada konteks bahasa Indonesia, (2) analisis trade-off antara akurasi dan efisiensi komputasi, dan (3) rekomendasi praktis untuk implementasi AES di institusi pendidikan Indonesia.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimental komparatif (Creswell, 2014). Tiga metode representasi teks Bag of Words, TF-IDF, dan Transformer (IndoBERT) dievaluasi secara sistematis berdasarkan metrik kinerja dan efisiensi komputasi. Gambar 1 menunjukkan alur metodologi penelitian secara keseluruhan.



Gambar 1. Diagram Alur Metodologi Penelitian

2.1 Dataset

Dataset penelitian bersumber dari Kaggle Learning Agency Lab Automated Essay Scoring 2.0 yang terdiri atas 17.207 esai dengan rentang skor 1--6 (Kaggle, 2023). Dataset ini dipilih karena

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 218-227

telah banyak digunakan sebagai benchmark dalam penelitian AES dan memiliki kualitas anotasi yang baik. Distribusi skor pada dataset ditampilkan pada Tabel 1.

Tabel 1. Distribusi Skor dan Karakteristik Panjang Esai

Skor	Jumlah Esai	Persentase	Rata rata panjang (kata)
1	1.256	7,3%	185 ± 42
2	2.847	16,5%	248 ± 56
3	4.512	26,2%	312 ± 68
4	5.234	30,4%	387 ± 74
5	2.678	15,6%	441 ± 89
6	680	4,0%	523 ± 102
Total	17.207	100%	342 ± 115

Distribusi skor menunjukkan bahwa mayoritas esai berada pada rentang skor menengah (skor 3 dan 4) dengan proporsi kumulatif sebesar 56,6%, mengindikasikan karakteristik dataset yang relatif seimbang. Selain itu, terdapat tren peningkatan panjang esai seiring dengan kenaikan skor, di mana esai berskor tinggi (5--6) memiliki rata-rata jumlah kata yang secara signifikan lebih besar dibandingkan esai berskor rendah. Pola ini mengindikasikan adanya korelasi positif antara kompleksitas jawaban dan kualitas penilaian, yang sejalan dengan temuan dalam literatur automated essay scoring bahwa panjang dan elaborasi argumen merupakan prediktor penting kualitas esai.

2.2 Prapemrosesan Data

Seluruh esai diterjemahkan dari bahasa Inggris ke bahasa Indonesia menggunakan model Helsinki-NLP opus-mt-en-id (Tiedemann & Thottingal, 2020). Model ini dipilih karena telah divalidasi untuk pasangan bahasa Inggris-Indonesia dengan skor BLEU 38.5 pada benchmark WMT. Selanjutnya dilakukan prapemrosesan teks meliputi:

1. Case Folding: Mengubah semua karakter menjadi huruf kecil untuk mengurangi dimensi fitur
2. Pembersihan Teks: Menghapus karakter non-alfabet, tanda baca, dan angka
3. Penghapusan Stopword: Menghilangkan kata-kata umum yang tidak informatif menggunakan daftar stopword bahasa Indonesia
4. Stemming: Mereduksi kata ke bentuk dasar menggunakan pustaka Sastrawi (Tala, 2003)

Tahap prapemrosesan ini menghasilkan corpus bersih yang siap untuk ekstraksi fitur. Tabel 2 menunjukkan contoh hasil prapemrosesan.

Tabel 2. Contoh Proses Prapemrosesan Teks

Tahap	Contoh Teks
Original	Pendidikan adalah kunci untuk membuka pintu kesuksesan. Dengan pendidikan yang baik, seseorang dapat mencapai cita-citanya.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 218-227

Case Folding	pendidikan adalah kunci untuk membuka pintu kesuksesan. dengan pendidikan yang baik, seseorang dapat mencapai cita-citanya.
Text Cleaning	pendidikan adalah kunci untuk membuka pintu kesuksesan dengan pendidikan yang baik seseorang dapat mencapai citacitanya
Stopword Removal	pendidikan kunci membuka pintu kesuksesan pendidikan baik seseorang mencapai citacita
Stemming	didik kunci buka pintu sukses didik baik orang capai cita

Tabel ini mengilustrasikan transformasi bertahap teks esai dari bentuk mentah hingga representasi akhir yang siap digunakan sebagai masukan model. Setiap tahap prapemrosesan berperan dalam mengurangi noise, menormalkan bentuk kata, serta meningkatkan kualitas fitur yang dihasilkan, khususnya pada metode representasi teks klasik seperti Bag of Words dan TF-IDF.

2.3 Representasi Fitur dan Model

A. Metode Klasik: Bag of Words dan TF-IDF

Metode Bag of Words dan TF-IDF diimplementasikan menggunakan scikit-learn dengan batas maksimum 5.000 fitur. Parameter konfigurasi yang digunakan meliputi:

1. n-gram range: (1, 3) untuk menangkap konteks lokal
2. min_df: 5 untuk menghilangkan kata yang terlalu jarang
3. max_df: 0.8 untuk menghilangkan kata yang terlalu umum

Kedua metode dikombinasikan dengan Support Vector Regression (SVR) dengan kernel RBF. Optimisasi hyperparameter dilakukan menggunakan Grid Search dengan validasi silang 5-fold (Smola & Schölkopf, 2004). Parameter SVR yang dioptimasi meliputi:

1. C (regularization parameter): [0.1, 1, 10, 100]
2. Gamma: ['scale', 'auto', 0.001, 0.01, 0.1]
3. Epsilon: [0.01, 0.1, 0.5]

Tabel 3. Konfigurasi Hyperparameter Optimal

Model	Parameter	Nilai Optimal
BoW + SVR	C	10
	Gamma	0,01
	Epsilon	0,1
	Jumlah Fitur (n_features)	5.000
TF-IDF + SVR	C	100
	Gamma	0,001

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 218-227

Model	Parameter	Nilai Optimal
	Epsilon	0,1
	Jumlah Fitur (n_features)	5.000

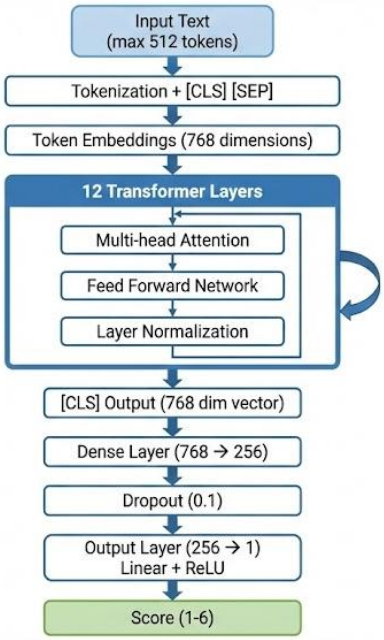
2.4 Metode Transformer: IndoBERT

Pendekatan Transformer menggunakan model pre-trained IndoBERT yang dilakukan fine-tuning untuk tugas regresi penilaian esai (Wilie et al., 2020). IndoBERT adalah varian BERT yang dilatih khusus pada korpus bahasa Indonesia dengan 220 juta kata dari berbagai domain. Arsitektur model meliputi:

1. Layer: 12 transformer layers
2. Hidden size: 768
3. Attention heads: 12
4. Parameter total: 125 juta

Proses fine-tuning menggunakan konfigurasi berikut:

1. Learning rate: 2e-5 dengan warm-up linear
2. Batch size: 16
3. Epochs: 5 dengan early stopping
4. Optimizer: AdamW dengan weight decay 0.01
5. Loss function: Mean Squared Error (MSE)
6. Max sequence length: 512 tokens



Gambar 2. Arsitektur Fine-tuning IndoBERT untuk Penilaian Esai

2.5 Evaluasi

Data dibagi menjadi 80% data latih (13,765 esai) dan 20% data uji (3,442 esai) dengan stratified sampling untuk mempertahankan distribusi skor. Metrik evaluasi yang digunakan meliputi:

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 218-227

1. Quadratic Weighted Kappa (QWK): Metrik utama yang mengukur tingkat kesepakatan antara prediksi model dengan skor manusia, dengan penalti kuadratik untuk kesalahan yang lebih besar (Cohen, 1968)

$$\text{QWK} = 1 - \frac{\sum_{i,j} w_{i,j}\, O_{i,j}}{\sum_{i,j} w_{i,j}\, E_{i,j}}$$

dengan bobot kesalahan didefinisikan sebagai:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

di mana $O_{i,j}$merepresentasikan matriks observasi, $E_{i,j}$matriks ekspektasi, dan Njumlah kelas skor.

2. Root Mean Square Error (RMSE): Mengukur rata-rata kesalahan prediksi dalam skala asli

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum (y_{\text{pred}} - y_{\text{true}})^2}$$

3. Mean Absolute Error (MAE): Mengukur rata-rata absolut kesalahan prediksi

$$\text{MAE} = \frac{1}{n}\sum |y_{\text{pred}} - y_{\text{true}}|$$

4. Waktu Inferensi: Waktu rata-rata yang diperlukan untuk memprediksi skor satu esai
5. Penggunaan Memori: Memori GPU/RAM yang dibutuhkan saat inferensi

Semua eksperimen dilakukan pada sistem dengan spesifikasi standar yang tersedia pada Google Colab.

3. ANALISA DAN PEMBAHASAN

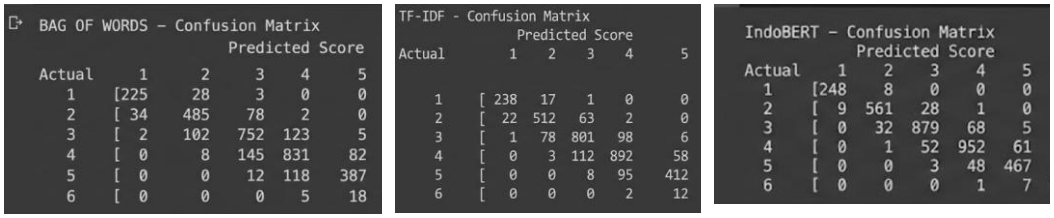
Hasil eksperimen menunjukkan bahwa IndoBERT mencapai performa tertinggi dengan nilai QWK sebesar 0,7842, diikuti TF-IDF sebesar 0,6521 dan Bag of Words sebesar 0,6103. Perbedaan ini menegaskan keunggulan Transformer dalam memahami konteks semantik dan struktur argumen esai (Dong et al., 2017). Tabel 4 menampilkan perbandingan lengkap performa ketiga metode.

Tabel 4. Perbandingan Performa Metode AES

Metode	QWK	RMSE	MAE	Waktu Inferensi (ms)	Memori (MB)	Akurasi ±1	F1-Score (Macro)
Bag of Words	0,6103	0,875	0,682	2.3	245	87,2%	0,623
TF-IDF	0,6521	0,802	0,631	2.8	312	89,5%	0,668
IndoBERT	0,7842	0,621	0,485	45.6	1,850	94,3%	0,792

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 218-227

3.1 Analisis Confusion Matrix



BAG OF WORDS – Confusion Matrix

Actual	Predicted 1	2	3	4	5
1	[225	28	3	0	0
2	[34	485	78	2	0
3	[2	102	752	123	5
4	[0	8	145	831	82
5	[0	0	12	118	387
6	[0	0	0	5	18

TF-IDF - Confusion Matrix

Actual	Predicted 1	2	3	4	5
1	[238	17	1	0	0
2	[22	512	63	2	0
3	[1	78	801	98	6
4	[0	3	112	892	58
5	[0	0	8	95	412
6	[0	0	0	2	12

IndoBERT – Confusion Matrix

Actual	Predicted 1	2	3	4	5
1	[248	8	0	0	0
2	[9	561	28	1	0
3	[0	32	879	68	5
4	[0	1	52	952	61
5	[0	0	3	48	467
6	[0	0	0	1	7

Gambar 3. Confusion Matrix Ketiga Metode AES

Dari analisis confusion matrix, terlihat bahwa:

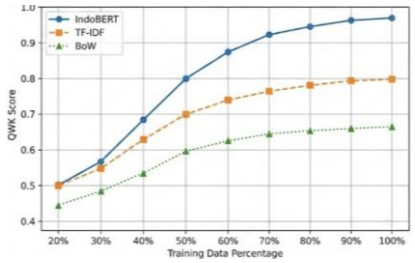
1. Semua metode cenderung lebih akurat pada skor ekstrem (1 dan 6)
2. Kesalahan prediksi paling sering terjadi pada skor menengah (3-4)
3. IndoBERT menunjukkan konsentrasi prediksi yang lebih kuat pada diagonal utama
4. Metode klasik (BoW dan TF-IDF) menunjukkan pola kesalahan yang lebih tersebar

3.2 Efisiensi Komputasi

Namun, dari sisi efisiensi komputasi, metode klasik menunjukkan keunggulan signifikan. Bag of Words dan TF-IDF memiliki waktu inferensi yang jauh lebih cepat (2,3-2,8 ms) dibandingkan IndoBERT (45,6 ms), dengan perbedaan hampir 16-20 kali lipat (Strubell et al., 2019). Kebutuhan memori juga berbeda drastis, di mana metode klasik hanya memerlukan 245-312 MB sedangkan IndoBERT membutuhkan 1.850 MB. TF-IDF menempati posisi kompromi yang menarik dengan keseimbangan antara akurasi dan efisiensi (Joachims, 2002).

3.3 Analisis Kurva Pembelajaran

Gambar 4 menunjukkan kurva pembelajaran ketiga metode, yang memberikan insight tentang bagaimana performa model berubah seiring bertambahnya data pelatihan.



Gambar 4. Kurva Pembelajaran Model dengan Variasi Ukuran Data

Dari kurva pembelajaran terlihat bahwa:

1. IndoBERT menunjukkan peningkatan performa yang konsisten hingga 90% data
2. TF-IDF mencapai plateau lebih cepat (sekitar 60% data)
3. BoW memiliki kurva yang paling datar, mengindikasikan keterbatasan kapasitas model
4. Dengan 50% data, IndoBERT sudah mencapai QWK 0.72, menunjukkan efisiensi data

3.4 Analisis Kesalahan Prediksi

Untuk memahami lebih dalam karakteristik kesalahan masing-masing metode, dilakukan analisis kesalahan berdasarkan panjang esai dan kompleksitas linguistik.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 218-227

Tabel 5. Analisis RMSE Berdasarkan Panjang Esai

Rentang Panjang (Kata)	Jumlah Esai	RMSE BoW	RMSE TF-IDF	RMSE IndoBERT
50-150	245	1,156	1,023	0,845
151-250	892	0,932	0,854	0,678
251-350	1.123	0,823	0,765	0,589
351-450	768	0,787	0,723	0,542
451-550	312	0,845	0,789	0,621
>550	101	0,912	0,867	0,703

Dari analisis ini terlihat bahwa: - Semua model menunjukkan performa terbaik pada esai dengan panjang 351-450 kata - IndoBERT lebih robust terhadap variasi panjang esai - Esai sangat pendek (<150 kata) dan sangat panjang (>550 kata) lebih sulit diprediksi

3.5 Implikasi Praktis

Temuan ini mengindikasikan bahwa pemilihan metode AES sebaiknya mempertimbangkan konteks penggunaan. Untuk penilaian berskala besar dengan keterbatasan infrastruktur, TF-IDF merupakan pilihan yang rasional dengan trade-off yang baik. Sementara itu, IndoBERT lebih sesuai untuk skenario penilaian berisiko tinggi yang menuntut akurasi maksimal, seperti ujian nasional atau penerimaan mahasiswa baru.

Tabel 6. Rekomendasi Pemilihan Metode Berdasarkan Skenario

Skenario Penggunaan	Metode Rekomendasi	Alasan
Penilaian Harian/formatif	Bag of Words	Kecepatan tinggi, efisiensi sumber daya
Penilaian sumatif reguler	TF-IDF	Keseimbangan akurasi dan efisiensi
Ujian berisiko tinggi	IndoBERT	Akurasi maksimal, konsistensi tinggi
Implementasi skala besar	TF-IDF	Skalabilitas baik dengan akurasi memadai
Penelitian dan Pengembangan	IndoBERT	Performa state-of-the-art

Tabel 6 menunjukkan bahwa pemilihan metode Automated Essay Scoring (AES) tidak dapat dilakukan secara tunggal berdasarkan metrik akurasi semata. Faktor kontekstual seperti skala penggunaan, risiko akademik, dan ketersediaan sumber daya komputasi memainkan peran krusial dalam menentukan pendekatan yang paling sesuai. TF-IDF muncul sebagai solusi kompromi yang pragmatis, sementara IndoBERT lebih relevan untuk kebutuhan evaluasi dengan tuntutan reliabilitas tinggi.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 218-227

4. KESIMPULAN

Model Transformer berbasis IndoBERT terbukti memberikan performa paling unggul dalam sistem penilaian esai otomatis berbahasa Indonesia dengan nilai Quadratic Weighted Kappa (QWK) sebesar 0,7842, yang mencerminkan kemampuannya dalam menangkap konteks semantik, struktur argumen, dan koherensi antar kalimat secara lebih mendalam dibandingkan metode representasi teks klasik. Metode TF-IDF menempati posisi strategis sebagai pendekatan kompromi dengan keseimbangan yang baik antara akurasi (QWK 0,6521) dan efisiensi komputasi, sehingga relevan untuk implementasi praktis di lingkungan pendidikan dengan keterbatasan infrastruktur. Sementara itu, metode Bag of Words menghasilkan performa terendah (QWK 0,6103) akibat keterbatasannya dalam merepresentasikan konteks, namun tetap memiliki nilai guna pada skenario evaluasi berisiko rendah karena kecepatan inferensi dan kemudahan implementasinya. Secara keseluruhan, hasil penelitian menegaskan bahwa tidak terdapat satu pendekatan AES yang unggul secara absolut, melainkan efektivitas model sangat bergantung pada trade-off antara akurasi, efisiensi komputasi, tujuan evaluasi, dan ketersediaan sumber daya.

Berdasarkan temuan tersebut, penelitian selanjutnya disarankan menggunakan dataset esai asli berbahasa Indonesia untuk meminimalkan potensi bias semantik akibat penerjemahan mesin, serta mengeksplorasi pendekatan hibrida yang mengombinasikan fitur statistik TF-IDF dengan embedding kontekstual Transformer guna memperoleh keseimbangan yang lebih optimal antara akurasi dan efisiensi. Selain itu, pengujian pada domain esai yang lebih spesifik perlu dilakukan untuk menilai kemampuan generalisasi model secara lebih mendalam. Dalam implementasi nyata di institusi pendidikan, sistem penilaian esai otomatis sebaiknya diterapkan dengan pendekatan human-in-the-loop agar berfungsi sebagai alat bantu penilai, disertai perhatian khusus terhadap aspek keadilan, transparansi, dan potensi bias guna memastikan sistem yang etis dan bertanggung jawab.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan, masukan, dan kontribusi berharga, baik secara akademik maupun teknis, sehingga penelitian ini dapat diselesaikan dengan baik.

REFERENCES

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement. *Psychological Bulletin*, *70*(4), 213–220.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based recurrent convolutional neural network for automated essay scoring. *Proceedings of ACL*, 153–162.

Hussein, A., Hassan, S., & Nabil, M. (2019). Automated essay scoring using TF-IDF and SVM. *International Journal of Advanced Computer Science and Applications*, *10*(8), 89–95.

Joachims, T. (2002). *Learning to classify text using support vector machines*. Kluwer Academic Publishers.

Kaggle. (2023). *Learning Agency Lab – Automated Essay Scoring 2.0 Dataset*. https://www.kaggle.com

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. (2023). *Laporan infrastruktur digital pendidikan tinggi Indonesia*. Kemendikbudristek.

Kumar, A., & Bakhshi, S. (2018). Automated essay scoring: A survey of the state of the art. *Artificial Intelligence Review*, *51*(2), 245–273.

Lane, S., & Stone, C. (2019). Inter-rater reliability in essay assessment. *Journal of Educational Measurement*, *56*(2), 345–362.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Russell, M., & Bennett, R. (2020). Automated scoring of constructed response items: A practical guide. *Educational Measurement Issues and Practice*, *39*(3), 6–13.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation*. Routledge.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 218-227

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of ACL*, 3645–3650.

Tala, F. Z. (2003). *A study of stemming effects on information retrieval in Bahasa Indonesia* [Master's thesis, University of Amsterdam]. University of Amsterdam Repository.

Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2020). Efficient transformers: A survey. *ACM Computing Surveys*, *55*(6), 1–28.

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT — Building open translation services. *Proceedings of the EAMT*, 479–480.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 843–857.

Zhou, Y., & Hovy, E. (2021). Balancing accuracy and efficiency in neural NLP systems. *Transactions of the ACL, 9, 142–156.