GAUSS : Journal of Mathematics, Statistics, and Mathematics Education
*e-ISSN: **3124-3681***
journal.ininnawaparaedu.com
Vol 02, No 01, 87-97

Analisis Matematis Metode *Automatic Short Answer* Scoring Bahasa Indonesia: Dataset, Tantangan, dan Arah Penelitian Masa Depan

[1]Nur Fadilah 1*
[1] Program Studi Matematika, Universitas Negeri Makassar, Makassar
*Email: nurfadilah@unm.ac.id

ABSTRAK
Automatic Short Answer Scoring (ASAS) merupakan salah satu penerapan *Natural Language Processing* (NLP) yang digunakan untuk melakukan penilaian otomatis terhadap jawaban uraian singkat. Pengembangan ASAS menjadi penting karena proses penilaian manual sering kali membutuhkan waktu yang lama dan rentan terhadap subjektivitas penilai. Meskipun penelitian ASAS Bahasa Indonesia telah berkembang dalam beberapa tahun terakhir, kajian yang mengintegrasikan analisis dataset, perkembangan metode, landasan matematis, tantangan, dan arah penelitian masa depan masih relatif terbatas. Oleh karena itu, penelitian ini bertujuan untuk melakukan *Systematic Literature Review* (SLR) terhadap perkembangan ASAS Bahasa Indonesia dengan fokus pada analisis matematis metode yang digunakan. Proses kajian dilakukan melalui pencarian dan seleksi literatur pada beberapa basis data ilmiah yang relevan dengan kriteria inklusi tertentu. Data yang diperoleh dianalisis secara deskriptif dan komparatif berdasarkan karakteristik dataset, metode yang digunakan, formulasi matematis, serta performa yang dilaporkan. Hasil kajian menunjukkan bahwa penelitian ASAS Bahasa Indonesia masih didominasi oleh penggunaan dataset UKARA, PISA Indonesia, Rahutomo, dan dataset domain khusus. Dari sisi metodologi, terjadi pergeseran dari pendekatan berbasis kemiripan teks seperti TF-IDF, *Vector Space Model* (VSM), *Cosine Similarity*, dan *Latent Semantic Analysis* (LSA) menuju pendekatan berbasis *deep learning* dan Transformer. Model berbasis BiLSTM dan FastText menunjukkan peningkatan performa dibandingkan metode konvensional, sedangkan pendekatan Transformer menghasilkan kemampuan pemahaman semantik yang lebih baik dan performa yang lebih stabil. Kajian ini juga mengidentifikasi tantangan utama berupa keterbatasan dataset, *semantic noise*, *overfitting*, dan variasi bahasa nonformal. Penelitian selanjutnya berpotensi mengembangkan model berbasis *Large Language Models* (LLM), pendekatan hibrida, dan sistem penilaian yang mampu memberikan umpan balik otomatis kepada peserta didik.

Kata Kunci: ***Automatic Short Answer Scoring; Deep Learning; Natural Language Processing; Systematic Literature Review*; Transformer**

ABTRACT
Automatic Short Answer Scoring *(ASAS) is an application of* Natural Language Processing *(NLP) that enables the automatic assessment of short-answer responses. The development of ASAS has become increasingly important because manual grading is often time-consuming and susceptible to subjective judgments. Although research on Indonesian ASAS has grown significantly in recent years, comprehensive studies integrating dataset analysis, methodological development, mathematical foundations, challenges, and future research directions remain limited. Therefore, this study aims to conduct a* Systematic Literature Review *(SLR) on the development of Indonesian ASAS with a particular focus on the mathematical analysis of the underlying methods. The review process was carried out through literature searching and screening across several scientific databases using predefined inclusion criteria. The collected studies were analyzed descriptively and comparatively based on dataset characteristics, employed methods, mathematical formulations, and reported performance. The results indicate that Indonesian ASAS research is primarily based on the UKARA, PISA Indonesia, Rahutomo, and domain-specific datasets. From a methodological perspective, research has evolved from text similarity approaches such as TF-IDF,* Vector Space Model *(VSM),* Cosine Similarity, *and* Latent Semantic Analysis *(LSA) toward* deep learning *and Transformer-based approaches. BiLSTM and FastText models demonstrate improved performance compared to conventional techniques, while Transformer-based methods provide stronger semantic understanding and more robust scoring performance. The review also identifies several key challenges, including limited datasets,*

semantic noise, overfitting, *and the diversity of informal language usage. Future research opportunities include the integration of* Large Language Models *(LLMs), hybrid scoring approaches, and automated feedback systems to support more effective learning assessment.*

Keywords: Automatic Short Answer Scoring; Deep Learning; Natural Language Processing; Systematic Literature Review; Transformer

1. PENDAHULUAN

Evaluasi merupakan salah satu komponen penting dalam proses pembelajaran karena berfungsi untuk mengukur tingkat pencapaian kompetensi peserta didik. Selain digunakan untuk menentukan hasil belajar, evaluasi juga berperan sebagai dasar dalam pengambilan keputusan terkait perbaikan strategi pembelajaran. Di antara berbagai bentuk evaluasi yang digunakan dalam pendidikan, soal uraian singkat (*short answer*) dan esai memiliki kemampuan yang lebih baik dalam mengukur pemahaman konseptual, kemampuan analisis, serta keterampilan berpikir tingkat tinggi dibandingkan soal pilihan ganda. Melalui jawaban tertulis, peserta didik dituntut untuk mengorganisasi gagasan, menjelaskan konsep, dan menyusun argumen secara mandiri. Namun demikian, proses penilaian jawaban uraian secara manual memerlukan waktu yang relatif lama dan rentan terhadap subjektivitas penilai (Ayu & Nurjanah, 2025).

Perkembangan teknologi kecerdasan buatan (*Artificial Intelligence*) dan *Natural Language Processing* (NLP) mendorong lahirnya berbagai sistem penilaian otomatis yang mampu membantu proses evaluasi pembelajaran. Salah satu pendekatan yang berkembang pesat adalah *Automatic Short Answer Scoring* (ASAS), yaitu sistem yang dirancang untuk memberikan skor secara otomatis terhadap jawaban uraian singkat berdasarkan tingkat kesesuaian jawaban peserta didik dengan kriteria penilaian yang telah ditentukan (Hidayatulloh, 2018). Penggunaan ASAS tidak hanya mampu meningkatkan efisiensi proses penilaian, tetapi juga berpotensi menghasilkan penilaian yang lebih konsisten dibandingkan proses koreksi manual yang dipengaruhi oleh faktor kelelahan dan subjektivitas manusia (Ayu & Nurjanah, 2025; Wicaksono dkk., 2024).

Dalam perkembangannya, penelitian ASAS mengalami transformasi metodologis yang signifikan. Generasi awal sistem ASAS banyak memanfaatkan pendekatan berbasis kemiripan teks (*text similarity*) yang mengandalkan representasi matematis dokumen menggunakan *Vector Space Model* (VSM), *Term Frequency-Inverse Document Frequency* (TF-IDF), *Cosine Similarity*, serta *Latent Semantic Analysis* (LSA) untuk mengukur tingkat kesamaan antara jawaban siswa dan kunci jawaban (Febriyanto, 2019; Arfandy & Musdar, 2020; Hidayatulloh, 2018). Pendekatan ini relatif sederhana dan mudah diimplementasikan, tetapi memiliki keterbatasan dalam memahami makna semantik yang terkandung di dalam teks.

Seiring berkembangnya kemampuan komputasi dan ketersediaan data, penelitian ASAS mulai mengadopsi pendekatan berbasis *machine learning* dan *deep learning*. Representasi kata menggunakan *word embedding* seperti FastText memungkinkan model memahami hubungan semantik antarkata secara lebih baik dibandingkan metode berbasis frekuensi kata. Di sisi lain, model *Bidirectional Long Short-Term Memory* (BiLSTM) mampu mempelajari dependensi kontekstual dari

urutan kata dalam kalimat sehingga menghasilkan performa yang lebih baik pada berbagai tugas klasifikasi teks (Fadilah & Priyanta, 2022). Penelitian terbaru bahkan menunjukkan bahwa optimasi arsitektur BiLSTM melalui integrasi *Global Max Pooling* dan *Class Weighting* mampu meningkatkan kemampuan model dalam menangani data yang tidak seimbang serta memperbaiki performa klasifikasi pada dataset ASAS Bahasa Indonesia (Fadilah dkk., 2026a).

Perkembangan terkini ditandai dengan dominasi arsitektur Transformer dan model bahasa besar (*Large Language Models*). Berbeda dengan pendekatan sebelumnya yang mengandalkan pencocokan kata atau representasi statis, model berbasis Transformer memanfaatkan mekanisme *attention* untuk memahami hubungan kontekstual antar kata secara lebih mendalam. Pendekatan ini memungkinkan sistem mengenali kesetaraan makna meskipun jawaban siswa memiliki struktur kalimat yang berbeda dengan kunci jawaban. Wicaksono dkk. (2024) menunjukkan bahwa pendekatan *direct scoring* berbasis Transformer menghasilkan performa yang lebih stabil dibandingkan pendekatan berbasis kemiripan semantik, terutama pada skenario pengujian lintas topik (*cross-prompt evaluation*).

Meskipun berbagai metode telah dikembangkan, implementasi ASAS untuk Bahasa Indonesia masih menghadapi sejumlah tantangan. Ketersediaan dataset publik yang terbatas, ukuran data yang relatif kecil, variasi bahasa nonformal, penggunaan singkatan, kesalahan penulisan, serta keberagaman struktur kalimat menyebabkan proses penilaian otomatis menjadi lebih kompleks dibandingkan bahasa dengan sumber daya linguistik yang lebih matang (Hidayatulloh, 2018; Ayu & Nurjanah, 2025). Selain itu, penggunaan teknik augmentasi data yang tidak terkontrol berpotensi menghasilkan *semantic noise* yang dapat menurunkan kemampuan generalisasi model (Fadilah dkk., 2026). Tantangan lain yang juga sering ditemukan adalah risiko *overfitting* pada model modern dan inkonsistensi skor yang berasal dari penilai manusia sebagai sumber data pelatihan (Wicaksono dkk., 2024).

Berbagai penelitian mengenai ASAS Bahasa Indonesia telah dipublikasikan dalam beberapa tahun terakhir. Namun, sebagian besar penelitian berfokus pada pengembangan model tertentu tanpa memberikan analisis komprehensif mengenai landasan matematis yang mendasari setiap pendekatan. Padahal, pemahaman terhadap formulasi matematis suatu metode sangat penting untuk menjelaskan bagaimana sistem melakukan representasi teks, menghitung kemiripan jawaban, mengekstraksi fitur semantik, hingga menghasilkan prediksi skor akhir. Kajian yang mengintegrasikan aspek dataset, perkembangan metode, formulasi matematis, tantangan implementasi, dan arah penelitian masa depan masih relatif terbatas.

Berdasarkan kondisi tersebut, penelitian ini bertujuan untuk melakukan *Systematic Literature Review* terhadap perkembangan *Automatic Short Answer Scoring* Bahasa Indonesia dengan fokus pada analisis matematis metode yang digunakan. Kajian ini mencakup identifikasi karakteristik dataset yang tersedia, formulasi matematis dari berbagai metode yang digunakan, perkembangan pendekatan dari metode berbasis kemiripan teks hingga model berbasis Transformer, tantangan yang dihadapi dalam implementasi sistem, serta peluang penelitian yang dapat dikembangkan pada masa mendatang. Hasil penelitian diharapkan dapat memberikan pemahaman yang lebih komprehensif mengenai fondasi matematis

ASAS Bahasa Indonesia sekaligus menjadi referensi bagi pengembangan sistem penilaian otomatis yang lebih akurat, robust, dan adaptif.

2. METODE PENELITIAN

Penelitian ini menggunakan metode *Systematic Literature Review* (SLR) dengan mengacu pada pedoman *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA 2020) (Page dkk., 2021). Kajian difokuskan pada perkembangan penelitian *Automatic Short Answer Scoring* (ASAS) Bahasa Indonesia, meliputi karakteristik dataset, metode yang digunakan, tantangan pengembangan sistem, serta arah penelitian masa depan.

Pencarian literatur dilakukan pada Google Scholar, Scopus, Semantic Scholar, dan Garuda menggunakan kata kunci "*Automatic Short Answer Scoring*", "*Automated Essay Scoring*", "*Bahasa Indonesia*", "*Natural Language Processing*", "*Deep Learning*", dan "*Transformer*". Artikel yang dipilih merupakan publikasi tahun 2018–2026 yang tersedia dalam bentuk teks lengkap.

Proses seleksi literatur mengikuti tahapan PRISMA, yaitu *identification*, *screening*, *eligibility*, dan *inclusion*. Kriteria inklusi meliputi artikel yang membahas penilaian otomatis jawaban uraian, penggunaan metode NLP untuk penilaian jawaban singkat, serta penelitian yang melaporkan dataset, metode, atau hasil evaluasi sistem. Data yang diperoleh kemudian dianalisis secara deskriptif dan komparatif untuk memetakan perkembangan metode ASAS Bahasa Indonesia dari pendekatan berbasis kemiripan teks hingga model berbasis *deep learning* dan Transformer

3. HASIL DAN PEMBAHASAN

Hasil

Berdasarkan proses seleksi literatur yang dilakukan, diperoleh sejumlah artikel yang membahas pengembangan *Automatic Short Answer Scoring* (ASAS) Bahasa Indonesia menggunakan berbagai pendekatan, mulai dari metode berbasis kemiripan teks hingga model berbasis *deep learning* dan Transformer. Literatur yang dianalisis dipublikasikan pada rentang tahun 2018–2026 dan berfokus pada pengembangan dataset, metode penilaian otomatis, serta evaluasi performa sistem.

3.1 Dataset *Automatic Short Answer Scoring* Bahasa Indonesia

Hasil kajian menunjukkan bahwa penelitian ASAS Bahasa Indonesia masih menggunakan jumlah dataset yang relatif terbatas. Dataset yang paling banyak digunakan adalah UKARA, diikuti oleh dataset PISA Indonesia, Rahutomo, dan dataset kejuruan SMKN 1 Rao Selatan. Masing-masing dataset memiliki karakteristik yang berbeda, baik dari sisi ukuran data, domain, maupun skema penilaian yang digunakan.

Tabel 1. Karakteristik Dataset ASAS Bahasa Indonesia

Dataset	Karakteristik Utama	Referensi
UKARA A	Jawaban relatif homogen dengan skema klasifikasi biner.	Hidayatulloh (2018); Fadilah dkk. (2026a)
UKARA B	Memiliki variasi semantik dan struktur kalimat yang lebih tinggi	Fadilah dkk. (2026a)

	dibandingkan UKARA A.	
PISA Indonesia	Menggunakan skema penilaian dikotomis dan politomis.	Hidayatulloh (2018)
Rahutomo	Digunakan untuk evaluasi lintas topik (*cross-prompt evaluation*).	Wicaksono dkk. (2024)
SMKN 1 Rao Selatan	Berisi jawaban siswa pada bidang Teknik Komputer dan Jaringan (TKJ).	Ayu & Nurjanah (2025)

3.2 Metode yang Digunakan dalam Penelitian ASAS Bahasa Indonesia

Berdasarkan literatur yang dianalisis, metode yang digunakan dalam penelitian ASAS Bahasa Indonesia dapat dikelompokkan menjadi tiga kategori utama, yaitu metode berbasis kemiripan teks, metode berbasis *deep learning*, dan metode berbasis Transformer. Perkembangan metode menunjukkan adanya pergeseran dari pendekatan berbasis pencocokan kata menuju pendekatan yang mampu memahami konteks dan makna semantik secara lebih mendalam.

Tabel 2. Kelompok Metode ASAS Bahasa Indonesia

Kelompok Metode	Metode
Berbasis Kemiripan Teks	TF-IDF, VSM, Cosine Similarity, LSA, GAN-LCS
Berbasis *Deep Learning*	FastText, BiLSTM, BiLSTM + EDA
Berbasis Transformer	SBERT, Direct Scoring, Transformer-Based Scoring

3.3 Hasil Performa Metode ASAS

Beberapa penelitian melaporkan performa yang berbeda sesuai dengan metode dan dataset yang digunakan. Metode berbasis *deep learning* menunjukkan peningkatan performa dibandingkan pendekatan berbasis kemiripan teks, sedangkan model berbasis Transformer menghasilkan performa yang lebih stabil pada skenario evaluasi lintas topik. Hasil tersebut menunjukkan bahwa penelitian ASAS Bahasa Indonesia terus berkembang dengan memanfaatkan metode yang semakin kompleks dan mampu menangkap hubungan semantik antarkalimat secara lebih efektif.

Tabel 3. Ringkasan Performa Metode ASAS Bahasa Indonesia

Penelitian	Metode	Dataset	Hasil
Fadilah & Priyanta (2022)	BiLSTM + EDA	UKARA	Akurasi 85,07%
Fadilah & Priyanta (2022)	BiLSTM + Random Insertion	UKARA B	Akurasi 72,78%
Wicaksono dkk. (2024)	Direct Scoring Transformer	Rahutomo	Pearson Correlation 0,9504
Fadilah dkk. (2026a)	BiLSTM-FastText + GMP	UKARA A	Akurasi 93,49%

Pembahasan

3.4 Analisis Matematis Metode Berbasis Kemiripan Teks

Metode berbasis kemiripan teks merupakan pendekatan awal yang banyak digunakan dalam penelitian *Automatic Short Answer Scoring* (ASAS) Bahasa Indonesia. Pendekatan ini bekerja dengan mengubah jawaban siswa dan kunci jawaban ke dalam bentuk representasi numerik, kemudian menghitung tingkat kemiripan keduanya menggunakan ukuran matematis tertentu. Metode yang paling umum digunakan adalah *Term Frequency–Inverse Document Frequency* (TF-IDF), *Vector Space Model* (VSM), *Cosine Similarity*, *Latent Semantic Analysis* (LSA), dan *Longest Common Subsequence* (LCS) (Hidayatulloh, 2018; Arfandy & Musdar, 2020). Berikut adalah rumus pembobotan TF-IDF

$$TFIDF(t,d) = TF(t,d) \times \log\left(\frac{N}{df(t)}\right)$$

Keterangan:

- $TF(t,d)$menunjukkan frekuensi kemunculan kata $t$pada dokumen d.
- Nadalah jumlah dokumen dalam korpus.
- $df(t)$adalah jumlah dokumen yang mengandung kata t.

Persamaan TF-IDF digunakan untuk memberikan bobot yang lebih tinggi pada kata-kata yang dianggap penting dalam suatu dokumen. Dalam konteks ASAS, representasi ini digunakan untuk mengubah jawaban siswa dan kunci jawaban menjadi vektor numerik yang dapat dibandingkan secara matematis. Selanjutnya rumus *cosine similarity* adalah sebagai berikut:

$$CosSim(A,B) = \frac{A \cdot B}{\| A \| \| B \|}$$

Setelah dokumen direpresentasikan dalam bentuk vektor, tingkat kemiripan dihitung menggunakan *Cosine Similarity*. Nilai yang semakin mendekati 1 menunjukkan tingkat kesamaan yang semakin tinggi antara jawaban siswa dan kunci jawaban.

Pada pendekatan TF-IDF, setiap kata diberikan bobot berdasarkan frekuensi kemunculannya dalam dokumen dan tingkat kelangkaannya dalam korpus. Selanjutnya, tingkat kemiripan antara jawaban siswa dan kunci jawaban dihitung menggunakan *Cosine Similarity* yang mengukur kedekatan dua vektor pada ruang multidimensi. Pendekatan ini relatif sederhana dan efisien karena tidak memerlukan data pelatihan dalam jumlah besar.

Meskipun demikian, metode berbasis kemiripan teks memiliki keterbatasan dalam memahami hubungan semantik antar kata. Jawaban yang memiliki makna sama tetapi menggunakan sinonim atau struktur kalimat berbeda sering kali memperoleh nilai kemiripan yang rendah. Untuk mengatasi masalah tersebut, beberapa penelitian mengadopsi *Latent Semantic Analysis* (LSA) yang memanfaatkan dekomposisi matriks untuk menemukan hubungan semantik tersembunyi antar kata dan dokumen.

Hasil kajian menunjukkan bahwa metode berbasis kemiripan teks masih efektif digunakan pada dataset berukuran kecil dan domain yang relatif homogen. Namun, kemampuan metode ini dalam menangkap konteks bahasa yang kompleks masih terbatas sehingga performanya cenderung lebih rendah dibandingkan pendekatan berbasis *deep learning* dan Transformer yang berkembang pada penelitian-penelitian berikutnya.

3.5 Analisis Matematis Metode Berbasis Kemiripan Teks

Perkembangan *deep learning* membawa peningkatan kemampuan sistem *Automatic Short Answer Scoring* (ASAS) dalam memahami hubungan semantik antar kata dan kalimat. Berbeda dengan metode berbasis kemiripan teks, pendekatan ini memanfaatkan representasi vektor (*word embedding*) seperti FastText untuk menangkap makna kata berdasarkan konteks penggunaannya. FastText yang diperkenalkan oleh Bojanowski dkk. (2017) memiliki keunggulan dalam memanfaatkan informasi sub-kata (*subword information*) sehingga mampu merepresentasikan kata yang jarang muncul maupun variasi morfologi yang umum ditemukan pada Bahasa Indonesia.

Representasi vektor yang dihasilkan kemudian diproses menggunakan arsitektur *Bidirectional Long Short-Term Memory* (BiLSTM) untuk mempelajari hubungan kontekstual antar kata dalam suatu kalimat. BiLSTM bekerja dengan memproses urutan kata dari dua arah, yaitu maju (*forward*) dan mundur (*backward*), sehingga mampu menangkap informasi konteks yang lebih lengkap dibandingkan LSTM konvensional. Secara matematis, keluaran setiap unit LSTM direpresentasikan oleh persamaan berikut:

$$h_t = o_t \odot \tanh(c_t)$$

Pada persamaan tersebut, (h_t) menyatakan *hidden state* pada waktu ke-(t), (o_t) merupakan *output gate*, dan (c_t) adalah *cell state* yang berfungsi menyimpan informasi penting dari urutan kata sebelumnya. Mekanisme ini memungkinkan model mempertahankan informasi kontekstual yang relevan sehingga lebih efektif dalam memahami makna jawaban siswa. Fadilah dan Priyanta (2022) melaporkan bahwa kombinasi BiLSTM dan *Easy Data Augmentation* (EDA) menghasilkan akurasi sebesar 85,07% pada dataset UKARA. Teknik EDA yang diperkenalkan oleh Wei dan Zou (2019) meningkatkan keragaman data pelatihan melalui operasi sederhana seperti *synonym replacement*, *random insertion*, *random swap*, dan *random deletion*, sehingga membantu mengurangi dampak keterbatasan dataset.

Penelitian yang lebih baru menunjukkan peningkatan performa melalui integrasi FastText, BiLSTM, *Global Max Pooling*, dan *Class Weighting*. Pendekatan tersebut mencapai akurasi 93,49% pada dataset UKARA A (Fadilah dkk., 2026). Hasil ini menunjukkan bahwa metode berbasis *deep learning* lebih efektif dalam menangkap hubungan semantik dibandingkan metode berbasis kemiripan teks, meskipun masih memerlukan data pelatihan yang memadai untuk menghindari *overfitting*.

3.6 Analisis Matematis Metode Berbasis Transformer

Perkembangan terbaru dalam penelitian *Automatic Short Answer Scoring* (ASAS) ditandai dengan penggunaan model berbasis Transformer yang mampu

memahami konteks kalimat secara lebih komprehensif dibandingkan metode sebelumnya. Arsitektur Transformer pertama kali diperkenalkan oleh Vaswani dkk. (2017) dan menjadi fondasi bagi berbagai model bahasa modern. Berbeda dengan pendekatan berbasis kemiripan teks maupun *deep learning* konvensional, Transformer memanfaatkan mekanisme *attention* untuk mempelajari hubungan antar kata tanpa bergantung pada pemrosesan sekuensial. Mekanisme *attention* memungkinkan model menentukan tingkat kepentingan setiap kata terhadap kata lainnya dalam suatu kalimat. Dengan demikian, model tidak hanya memperhatikan kata yang muncul, tetapi juga hubungan kontekstual yang terbentuk antar kata. Secara matematis, mekanisme *Scaled Dot-Product Attention* pada Transformer dirumuskan sebagai berikut:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Pada persamaan tersebut, (Q) (*Query*), (K) (*Key*), dan (V) (*Value*) merupakan representasi vektor yang digunakan untuk menghitung tingkat perhatian antar kata, sedangkan (d_k) menyatakan dimensi vektor *key*. Mekanisme ini memungkinkan model menghasilkan representasi kontekstual yang lebih kaya dibandingkan metode berbasis kemiripan teks maupun arsitektur rekuren seperti BiLSTM.

Melalui mekanisme tersebut, model dapat mengenali kesamaan makna meskipun jawaban siswa menggunakan susunan kalimat atau kosakata yang berbeda dari kunci jawaban. Pendekatan ini memungkinkan sistem melakukan penilaian berdasarkan representasi semantik yang lebih mendalam sehingga lebih adaptif terhadap variasi jawaban yang umum ditemukan pada soal uraian singkat. Wicaksono dkk. (2024) menunjukkan bahwa pendekatan *direct scoring* berbasis Transformer menghasilkan korelasi Pearson sebesar 0,9504 terhadap penilaian manusia. Hasil tersebut menunjukkan bahwa model berbasis Transformer memiliki kemampuan yang lebih baik dalam memahami konteks dan makna jawaban dibandingkan pendekatan berbasis kemiripan teks maupun model *deep learning* generasi sebelumnya.

3.7 Tantangan Pengembangan ASAS Bahasa Indonesia

Meskipun performa metode *Automatic Short Answer Scoring* (ASAS) terus mengalami peningkatan, pengembangannya untuk Bahasa Indonesia masih menghadapi beberapa tantangan. Salah satu tantangan utama adalah keterbatasan dataset publik yang tersedia. Sebagian besar penelitian masih menggunakan dataset yang relatif kecil, seperti UKARA dan beberapa dataset domain tertentu, sehingga berpotensi membatasi kemampuan generalisasi model (Hidayatulloh, 2018; Fadilah dkk., 2026).

Tantangan berikutnya berkaitan dengan variasi bahasa yang digunakan oleh peserta didik. Penggunaan sinonim, singkatan, bahasa tidak baku, serta variasi struktur kalimat menyebabkan jawaban yang memiliki makna sama dapat ditulis dalam bentuk yang sangat berbeda. Kondisi ini menyulitkan metode berbasis kemiripan teks dan menuntut model yang mampu memahami representasi semantik secara lebih mendalam (Ayu & Nurjanah, 2025).

Selain itu, penggunaan teknik augmentasi data yang tidak terkontrol berpotensi menghasilkan *semantic noise*, yaitu perubahan makna yang tidak sesuai

dengan jawaban asli. Pada model berbasis *deep learning*, tantangan lain yang sering ditemukan adalah *overfitting*, terutama ketika jumlah data pelatihan terbatas. Di sisi lain, kualitas penilaian manusia yang digunakan sebagai acuan (*ground truth*) juga dapat memengaruhi performa sistem karena adanya perbedaan interpretasi antarpenilai (Wicaksono dkk., 2024).

3.8 Arah Penelitian Masa Depan

Hasil kajian menunjukkan bahwa penelitian *Automatic Short Answer Scoring* (ASAS) Bahasa Indonesia masih memiliki peluang pengembangan yang luas. Salah satu arah penelitian yang menjanjikan adalah pengembangan teknik augmentasi data yang mampu meningkatkan jumlah data pelatihan tanpa mengubah makna semantik jawaban. Pendekatan ini penting mengingat keterbatasan dataset masih menjadi salah satu hambatan utama dalam pengembangan model ASAS Bahasa Indonesia.

Selain itu, integrasi pendekatan berbasis kemiripan semantik dan *direct scoring* berpotensi menghasilkan sistem yang lebih robust. Pendekatan hibrida memungkinkan model memanfaatkan keunggulan pengukuran kesamaan jawaban sekaligus kemampuan prediksi skor secara langsung. Seiring berkembangnya model berbasis Transformer, pemanfaatan *Large Language Models* (LLM) juga menjadi peluang penelitian yang menarik karena memiliki kemampuan memahami konteks dan penalaran bahasa yang lebih baik dibandingkan model generasi sebelumnya.

Arah penelitian lainnya adalah pengembangan sistem ASAS yang tidak hanya memberikan skor, tetapi juga mampu menghasilkan umpan balik otomatis (*automated feedback*) untuk membantu proses pembelajaran. Dengan dukungan teknologi NLP modern, sistem penilaian di masa depan diharapkan tidak hanya berfungsi sebagai alat evaluasi, tetapi juga sebagai sarana pendukung pembelajaran yang adaptif, akurat, dan mudah diimplementasikan dalam lingkungan pendidikan Indonesia.

4. KESIMPULAN

Kesimpulan menjelaskan rangkuman dari penelitian yang menjawab segala Penelitian ini telah menyajikan kajian literatur sistematis mengenai *Automatic Short Answer Scoring* (ASAS) Bahasa Indonesia dengan fokus pada analisis dataset, perkembangan metode, tantangan pengembangan, dan arah penelitian masa depan. Hasil kajian menunjukkan bahwa penelitian ASAS Bahasa Indonesia mengalami perkembangan yang signifikan dalam beberapa tahun terakhir. Dari sisi dataset, penelitian masih didominasi oleh penggunaan UKARA, PISA Indonesia, Rahutomo, dan dataset domain khusus seperti SMKN 1 Rao Selatan. Meskipun dataset tersebut telah berkontribusi dalam mendorong perkembangan penelitian ASAS, ketersediaan data publik yang terbatas masih menjadi salah satu hambatan utama dalam pengembangan model yang memiliki kemampuan generalisasi tinggi.

Dari perspektif metodologis, perkembangan ASAS menunjukkan pergeseran dari pendekatan berbasis kemiripan teks menuju model yang mampu memahami representasi semantik secara lebih mendalam. Metode awal seperti TF-IDF, VSM, *Cosine Similarity*, dan LSA menawarkan solusi yang sederhana dan efisien, namun memiliki keterbatasan dalam menangani variasi bahasa dan hubungan semantik antar kata. Perkembangan selanjutnya ditandai dengan pemanfaatan FastText dan

BiLSTM yang mampu mempelajari konteks kalimat secara lebih baik melalui representasi vektor dan pembelajaran mendalam. Penelitian terbaru menunjukkan bahwa model berbasis Transformer dan *direct scoring* memiliki performa yang lebih unggul karena mampu memahami konteks dan makna jawaban secara lebih komprehensif.

Kajian ini juga mengidentifikasi beberapa tantangan yang masih dihadapi dalam pengembangan ASAS Bahasa Indonesia, antara lain keterbatasan dataset, variasi bahasa nonformal, risiko *semantic noise* akibat augmentasi data, *overfitting* pada model berbasis *deep learning*, serta inkonsistensi penilaian manusia sebagai sumber *ground truth*. Oleh karena itu, penelitian selanjutnya perlu difokuskan pada pengembangan dataset yang lebih besar dan beragam, pemanfaatan model berbasis Transformer dan *Large Language Models* (LLM), serta pengembangan sistem yang tidak hanya menghasilkan skor otomatis tetapi juga mampu memberikan umpan balik yang konstruktif bagi peserta didik. Dengan demikian, sistem ASAS di masa depan diharapkan dapat menjadi solusi evaluasi pembelajaran yang lebih akurat, adaptif, dan relevan dengan kebutuhan pendidikan di Indonesia.

DAFTAR PUSTAKA

Arfandy, A., & Musdar, I. A. (2020). Automated Essay Scoring Menggunakan Cosine Similarity dan TF-IDF. Jurnal Karya Ilmiah, 22(2), 107-118. https://ejurnal.ubharajaya.ac.id/index.php/JKI/article/download/1684/1311

Ayu, R., & Nurjanah, D. (2025). Correlation Between Automatic Short Answer Scoring and Manual Scoring by Teacher on Indonesian Assessments. JTP - Jurnal Teknologi Pendidikan, 27(2), 751–764. https://journal.unj.ac.id/unj/index.php/jtp/article/view/48378/22108

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

Fadilah, N., & Priyanta, S. (2022). Automatic Essay Scoring Using Data Augmentation in Bahasa Indonesia. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 16(4), 401–410. https://journal.ugm.ac.id/ijccs/article/view/76396

Fadilah, N., Putra, B. A., & Pratama, M. I. (2026). Optimasi Model BiLSTM Berbasis FastText pada Data Augmentasi Semantik IndoBERT untuk Klasifikasi Teks Bahasa Indonesia. PISCES: Journal of Progressive Information, Security, Computer and Embedded System, 4(1), 1-10. https://journal.diginus.id/PISCES/article/view/1249/525

Fateen, M., Wang, B., & Mine, T. (2024). Beyond Scores: A Modular RAG-Based System for Automatic Short Answer Scoring With Feedback. IEEE Access, 12, 1249-1260. https://ishikawalab.ynu.ac.jp/ieee/access/index.html

Febriyanto, F. (2019). Sistem Penilaian Otomatis Jawaban Esai Dengan Menggunakan Metode Vector Space Model Pada Beberapa Perkuliahan Di STMIC Indonesia Banjarmasin. Jurnal Teknologi Informasi, 14(1), 53-68. http://jurnal.stkippersada.ac.id/jurnal/index.php/jutech/article/view/1273/0

Hidayatulloh. (2018). UKARA: A Fast and Simple Automatic Short Answer Scoring System for Bahasa Indonesia. Proceedings of the International Conference on Educational Assessment and Policy (ICEAP), 2(1), 48–53. https://pdfs.semanticscholar.org/0edf/00084126dea1d286fdf928dea691887fab79.pdf

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Prastowo, B. N., dkk. (2024). Stop-Word Exclusion and Reference Answer Comparison in Transformer-Based Automated Essay Scoring in Higher Education. Jurnal Nasional Teknologi dan Sistem Informasi (JANAPATI), 13(1), 103823. https://ejournal.undiksha.ac.id/index.php/janapati/article/view/103823

Sari, Y., dkk. (2019). A Platform View of Automatic Short Answer Scoring System. Proceedings of the International Conference on Educational Assessment and Policy (ICEAP), 3(1), 264-270. https://sinta.kemdiktisaintek.go.id/authors/profile/6021610/?view=googlescholar

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 6383–6389. https://doi.org/10.18653/v1/D19-1670

Wicaksono, B., Rasim, & Wihardi, Y. (2024). Analisis Komparatif Pendekatan Direct Scoring dan Similarity-Based Scoring pada Automatic Short Answer Scoring Berbahasa Indonesia Menggunakan Dataset Rahutomo. Brilliance: Research of Artificial Intelligence, 4(1), 6275. https://itscience-indexing.com/jurnal/index.php/brilliance/article/view/6275