PREDIKSI NILAI JAWABAN ESAI DENGAN MENGGUNAKAN *FASTTEXT* DAN ALGORITMA *LONG SHORT- TERM MEMORY (LSTM)*

TESIS

Oleh:
SEFTI AGUSTINI
NIM. 2206052200003

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026

PREDIKSI NILAI JAWABAN ESAI DENGAN MENGGUNAKAN *FASTTEXT* DAN ALGORITMA *LONG SHORT- TERM MEMORY (LSTM)*

TESIS

Diajukan kepada:
Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan Dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh:
SEFTI AGUSTINI
NIM. 2206052200003

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

PREDIKSI NILAI JAWABAN ESAI DENGAN MENGGUNAKAN *FASTTEXT* DAN ALGORITMA *LONG SHORT- TERM MEMORY (LSTM)*

TESIS

Oleh:
SEFTI AGUSTINI
NIM. 2206052200003

Telah diperiksa dan disetujui untuk diuji
Tanggal: 25 Juni 2026

Pembimbing I,

Dr. M. Ainul Yaqin, M.Kom
NIP. 19761013 200604 1 004

Pembimbing II,

Prof. Dr. Suhartono, S.Si., M.Kom
NIP 196805192 00312 1 001

Mengetahui,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Prof. Dr. Ir. Muhammad Faisal, S.Kom., M.T.
NIP. 19740510 200501 1 007

PREDIKSI NILAI JAWABAN ESAI DENGAN MENGGUNAKAN *FASTTEXT* DAN ALGORITMA *LONG SHORT- TERM MEMORY (LSTM)*

TESIS

Oleh:
SEFTI AGUSTINI
NIM. 2206052200003

Telah Dipertahankan di Depan Dewan Penguji
Dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
untuk Memperoleh Gelar Magister Komputer (M.Kom)
Tanggal: 25 Juni 2026

Susunan Dewan Penguji		**Tanda Tangan**
Penguji I	: Prof. Dr. Ir. Muhammad Faisal, S.Kom., MT NIP. 19740510 200501 1 007	( )
Penguji II	: Dr. Totok Chamidy, S.T., M.Kom NIP. 19691222 200604 1 001	( )
Pembimbing I	: Dr. M. Ainul Yaqin, M.Kom NIP. 19761013 200604 1 004	( )
Pembimbing II	: Prof. Dr. Suhartono, S.Si., M.Kom NIP 196805192 00312 1 001	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Prof. Dr. Ir. Muhammad Faisal, S.Kom., M.T.
NIP. 19740510 200501 1 007

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan dibawah ini:

Nama	: Sefti Agustini
NIM	: 2206052200003
Program Studi	: Magister Informatika
Fakultas	: Sains dan Teknologi
Judul Thesis	: "Prediksi Nilai Jawaban Esai dengan Menggunakan *Fasttext* dan Algoritma *Long Short- Term Memory (*LSTM*)"*

Menyatakan dengan sebenarnya bahwa Thesis yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan Thesis ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 25 Juni 2026

Yang membuat pernyataan,

Sefti Agustini

NIM. 2206052200003

HALAMAN MOTTO

“Allah tidak membebani seseorang melainkan sesuai dengan kesanggupannya.”
(QS. Al- Baqarah: 286)

“Jangan menyerah pada keadaan, karena setiap langkah kecil adalah bagian dari jalan menuju keberhasilan.”

HALAMAN PERSEMBAHAN

ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَالَمِينَ

Puji syukur atas kehadirat Allah SWT
Shalawat serta salam kepada Rasulullah SAW

Dengan segenap hati, penulis mempersembahkan sebuah karya ini kepada:

Suami tercinta Anton Apriansah, S.E., M.M. yang selalu menjadi penyemangat, dan selalu memberikan do'a dukungan, serta motivasi yang tidak terhingga.

Seluruh orang tua penulis tercinta, Bapak Wajdi dan Almh Ibu Helda yang selalu memberikan do'a, dukungan, serta motivasi.

Dosen pembimbing Dr. M. Ainul Yaqin, M.Kom. dan Prof. Dr. Suhartono, S.Si., M.Kom. yang telah membimbing penelitian ini dengan memberikan banyak pengarahan dan pengalaman yang berharga.

Segenap sivitas akademika Program Studi Magister Informatika, terutama seluruh dosen, terima kasih atas segenap ilmu dan bimbingannya.

Seluruh rekan-rekan mahasiswa Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Sahabat penulis Lilik Mauludiyah, S.Pd., M.Psi., Andin WAhyuning Putri, S.Pd., dan Fatimah, M.Pd. yang selalu menjadi penyemangat, dan selalu memberikan do'a dukungan, serta motivasi.

Penulis ucapkan "*jazakumullah khairan katsiiraa"*. Semoga selalu diridhoi Allah SWT. Aamiin Ya Rabbal 'Alamiin.

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Puji syukur penulis panjatkan ke hadirat Allah Swt. yang telah melimpahkan rahmat, taufik, serta hidayah-Nya sehingga penulis dapat menyelesaikan penyusunan tesis yang berjudul "*Prediksi Nilai Jawaban Esai dengan Menggunakan FastText dan Algoritma LSTM*" dengan baik. Selawat serta salam semoga senantiasa tercurahkan kepada junjungan kita Nabi Muhammad Saw., keluarga, sahabat, serta seluruh pengikutnya hingga akhir zaman.

Tesis ini disusun sebagai salah satu syarat untuk memperoleh gelar Magister Komputer (M.Kom) pada Program Studi Magister Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Penulis menyadari bahwa penyusunan tesis ini tidak terlepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, dengan segala kerendahan hati penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Prof. Dr. Hj. Ilfi Nur Diana, M.Si, selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Dr. H. Agus Mulyono, S.Pd., M.Kes, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Prof. Dr. Ir. Muhammad Faisal, S.Kom., M.T, selaku Ketua Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Dr. M. Ainul Yaqin, M.Kom, selaku Dosen Pembimbing I yang dengan penuh kesabaran telah meluangkan waktu untuk memberikan bimbingan, arahan, dan masukan yang sangat berharga selama proses penyusunan tesis ini.

5. Prof. Dr. Suhartono, S.Si., M.Kom, selaku Dosen Pembimbing II yang senantiasa memberikan bimbingan, saran, serta motivasi yang membangun hingga tesis ini dapat terselesaikan.
6. Seluruh dosen dan tenaga kependidikan Program Studi Magister Informatika Fakultas Sains dan Teknologi yang telah memberikan ilmu, wawasan, dan pelayanan selama masa studi.
7. Kepala Sekolah, guru mata pelajaran Bahasa Inggris, beserta siswa SMP Negeri 6 Singosari yang telah memberikan izin serta membantu penyediaan data penelitian yang menjadi dasar penyusunan tesis ini.
8. Kedua orang tua serta seluruh keluarga tercinta yang senantiasa memberikan doa, kasih sayang, dan dukungan yang tiada henti.
9. Rekan-rekan mahasiswa Program Studi Magister Informatika dan semua pihak yang tidak dapat penulis sebutkan satu per satu, yang telah memberikan bantuan dan semangat selama penyusunan tesis ini.

Semoga segala bantuan, bimbingan, dan dukungan yang telah diberikan mendapatkan balasan yang berlipat dari Allah Swt. Penulis menyadari bahwa tesis ini masih jauh dari kesempurnaan, sehingga kritik dan saran yang membangun sangat penulis harapkan demi perbaikan di masa mendatang. Akhir kata, penulis berharap semoga tesis ini dapat memberikan manfaat bagi pembaca serta bagi perkembangan ilmu pengetahuan, khususnya di bidang pemrosesan bahasa alami dan penilaian esai otomatis.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Malang, Juli 2026

Penulis,

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

ABSTRAK

Agustini, Sefti. 2026. *Prediksi Nilai Jawaban Esai dengan Menggunakan FastText dan Algoritma LSTM*. Tesis. Program Studi Magister Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. M. Ainul Yaqin, M.Kom; (II) Prof. Dr. Suhartono, S.Si., M.Kom.

Penilaian jawaban esai secara manual menuntut waktu dan tenaga yang besar, rentan terhadap subjektivitas dan inkonsistensi antar-penilai, serta memperlambat umpan balik kepada siswa. Penelitian ini bertujuan membangun model penilaian esai otomatis (Automated Essay Scoring) yang menggabungkan representasi teks FastText dengan algoritma Long Short-Term Memory (LSTM) untuk menghasilkan prediksi skor yang otomatis, akurat, konsisten, dan efisien. Data yang digunakan berupa 500 respons esai Bahasa Inggris dari 100 siswa SMP Negeri 6 Singosari yang masing-masing mengerjakan lima butir soal, dengan skor guru berskala 0–100 sebagai target regresi. Teks soal dan jawaban digabungkan menggunakan token khusus <startanswer> dan <endanswer>, kemudian direpresentasikan dengan FastText berdimensi 300 yang dilatih ulang pada korpus soal–jawaban sehingga tahan terhadap kata di luar kosakata (out-of-vocabulary). Dua arsitektur, yaitu LSTM dan Bidirectional LSTM (Bi-LSTM), dilatih dan dibandingkan setelah penalaan hiperparameter sistematis melalui 38 skenario, dengan konfigurasi optimal berupa 128 unit hidden, learning rate 0,001, batch size 32, dan dropout rate 0,3. Evaluasi kinerja menggunakan Mean Absolute Percentage Error (MAPE). Hasil pengujian menunjukkan Bi-LSTM memperoleh MAPE 7,2696% (kategori “Sangat Baik”, kurang dari 10%) dengan MAE 5,5958 dan RMSE 7,4171, sedikit mengungguli LSTM (MAPE 7,3747%); sebanyak 78,67% prediksi memiliki APE di bawah 10%. Model bekerja optimal pada jawaban panjang dan jawaban dengan skor di atas KKM, namun akurasinya menurun pada jawaban di bawah KKM akibat ketidakseimbangan kelas dan pada soal yang belum pernah dilihat saat pelatihan. Penelitian ini membuktikan bahwa pemaduan representasi subword FastText dengan pemodelan konteks sekuensial LSTM efektif untuk penilaian esai otomatis.

Kata Kunci: penilaian esai otomatis, FastText, LSTM, Bidirectional LSTM, MAPE

ABSTRACT

Agustini, Sefti. 2026. Essay Answer Score Prediction Using FastText and the LSTM Algorithm. Thesis. Master Program in Informatics, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Advisors: (I) Dr. M. Ainul Yaqin, M.Kom; (II) Prof. Dr. Suhartono, S.Si., M.Kom.

Manual essay scoring is time-consuming and labor-intensive, prone to subjectivity and inter-rater inconsistency, and delays feedback to students. This study aims to develop an Automated Essay Scoring model that combines FastText text representation with the Long Short-Term Memory (LSTM) algorithm to produce automatic, accurate, consistent, and efficient score predictions. The dataset comprises 500 English essay responses from 100 junior high school students of SMP Negeri 6 Singosari, each answering five essay questions, with teacher scores on a 0–100 scale as the regression target. The question and answer texts are concatenated using the special tokens <startanswer> and <endanswer>, then represented with a 300-dimensional FastText model retrained on the question–answer corpus so that it is robust to out-of-vocabulary words. Two architectures, LSTM and Bidirectional LSTM (Bi-LSTM), were trained and compared after systematic hyperparameter tuning across 38 scenarios, yielding an optimal configuration of 128 hidden units, a learning rate of 0.001, a batch size of 32, and a dropout rate of 0.3. Performance was evaluated using the Mean Absolute Percentage Error (MAPE). The results show that Bi-LSTM achieved a MAPE of 7.2696% (“Very Good” category, below 10%) with an MAE of 5.5958 and an RMSE of 7.4171, slightly outperforming LSTM (MAPE 7.3747%); 78.67% of predictions had an APE below 10%. The model performed best on long answers and above-passing-grade scores, but its accuracy declined on below-passing-grade answers due to class imbalance and on questions unseen during training. This study demonstrates that combining FastText subword representation with LSTM sequential context modeling is effective for automated essay scoring.

Keywords: automated essay scoring, FastText, LSTM, Bidirectional LSTM, MAPE

مستخلص البحث

أغوستيني، سفتي. 2026. التنبؤ بدرجات إجابات المقالات باستخدام FastText وخوارزمية LSTM. رسالة ماجستير. برنامج ماجستير المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية بمالانج. المشرف الأول: الدكتور محمد عين اليقين الماجستير؛ المشرف الثاني: الأستاذ الدكتور سوهارتونو الماجستير.

إن التصحيح اليدوي لإجابات المقالات يستغرق وقتًا وجهدًا كبيرين، وهو عرضة للذاتية وعدم الاتساق بين المصحّحين، كما يؤخّر التغذية الراجعة للطلاب. يهدف هذا البحث إلى بناء نموذج للتقييم الآلي للمقالات (Automated Essay Scoring) يجمع بين تمثيل النص FastText وخوارزمية الذاكرة الطويلة قصيرة المدى LSTM لإنتاج تنبؤات بالدرجات تتّسم بالآلية والدقة والاتساق والكفاءة. تتكوّن بيانات البحث من 500 إجابة مقالية باللغة الإنجليزية لـ 100 طالب في المدرسة المتوسطة الحكومية السادسة بسنجوساري، أجاب كل منهم عن خمسة أسئلة، مع اعتماد درجات المعلّم على مقياس من 0 إلى 100 هدفًا انحداريًا. تُدمج نصوص الأسئلة والإجابات باستخدام رمزين خاصّين هما startanswer وendanswer، ثم تُمثّل باستخدام نموذج FastText ذي 300 بُعد أُعيد تدريبه على مدوّنة الأسئلة والإجابات ليكون قادرًا على معالجة الكلمات خارج المفردات. وقد دُرّبت بنيتان هما LSTM وثنائية الاتجاه Bi-LSTM وقُورن بينهما بعد ضبط منهجي للمعاملات الفائقة عبر 38 سيناريو، وكان التكوين الأمثل: 128 وحدة مخفية، ومعدّل تعلّم 0.001، وحجم دفعة 32، ومعدّل إسقاط 0.3. واستُخدم متوسّط النسبة المئوية للخطأ المطلق (MAPE) في التقييم. أظهرت النتائج أن نموذج Bi-LSTM حقّق قيمة MAPE بلغت 7.2696% (ضمن فئة «ممتاز»، أقل من 10%) بقيمة MAE مقدارها 5.5958 وRMSE مقدارها 7.4171، متفوّقًا تفوّقًا طفيفًا على LSTM الذي بلغت قيمة MAPE فيه 7.3747%؛ كما أن 78.67% من التنبؤات كان خطؤها النسبي المطلق أقل من 10%. ويعمل النموذج على النحو الأمثل مع الإجابات الطويلة والدرجات التي تفوق حدّ النجاح، غير أن دقّته تنخفض مع الإجابات دون حدّ النجاح ومع الأسئلة التي لم تُعرض أثناء التدريب. ويُثبت هذا البحث أن الجمع بين تمثيل FastText على مستوى الوحدات الجزئية (subword) والنمذجة التسلسلية بخوارزمية LSTM فعّالٌ في التقييم الآلي لإجابات المقالات.

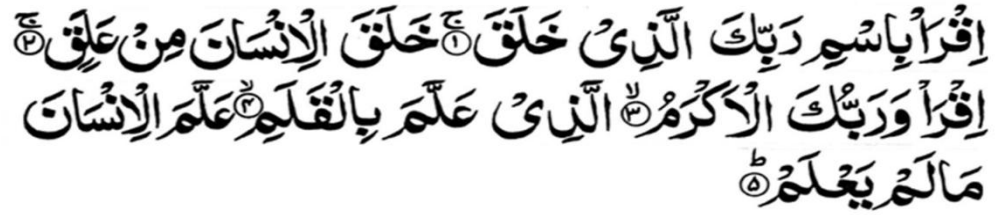
الكلمات المفتاحية: التقييم الآلي للمقالات، MAPE، Bi-LSTM، LSTM، FastText.

BAB I

PENDAHULUAN

1.1 Latar Belakang

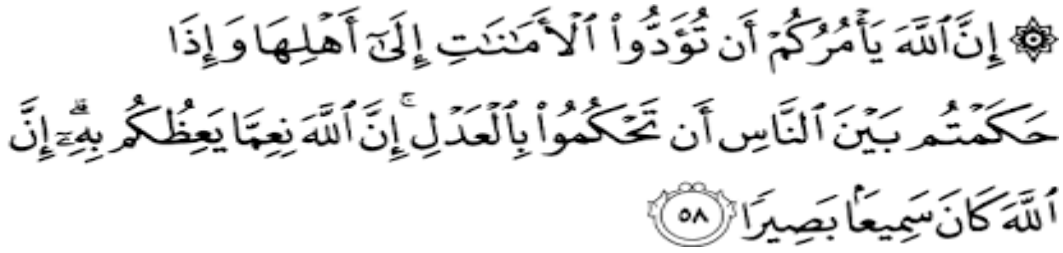
Dalam dunia pendidikan, salah satu bentuk evaluasi pembelajaran dilakukan dengan cara menilai hasil pekerjaan siswa (Maulani et al., 2024). Dari hasil pekerjaan tersebut akan diketahui kepahaman siswa terhadap materi yang sudah diberikan dan juga dapat dijadikan sebagai bahan evaluasi dalam sistem pembelajaran yang telah dilakukan. Dalam menilai pekerjaan siswa terutama untuk jawaban esai membutuhkan waktu ekstra jika dibandingkan dengan jawaban pilihan ganda karena jawaban esai memiliki beragam jawaban sesuai dengan keterampilan berpikir kritis, kemampuan berbahasa, mengorganisasi ide, serta pemahaman konsep siswa secara mendalam dan guru dalam mengoreksi harus dengan teliti mencocokkan dengan kunci jawaban yang sudah disediakan. Subjektivitas dalam menilai jawaban juga sering dilakukan karena berbagai kepentingan selain itu ketidakkonsistenan guru dalam menilai ketika jumlah jawaban yang akan nilai berada pada kelas besar sehingga memakan waktu lama yang pada akhirnya akan memperlambat umpan balik kepada siswa. Umpan balik sangat mempengaruhi proses pembelajaran siswa (Maulani et al., 2024), sesuai dengan anjuran dalam agama islam tentang kewajiban menuntut ilmu bagi setiap umat muslim yang tersurat dalam surat Al Alaq (69) ayat 1-5 :


اِقْرَاْ بِاسْمِ رَبِّكَ الَّذِيْ خَلَقَ ۚ﴿١﴾ خَلَقَ الْاِنْسَانَ مِنْ عَلَقٍ ۚ﴿٢﴾
اِقْرَاْ وَرَبُّكَ الْاَكْرَمُ ۙ﴿٣﴾ الَّذِيْ عَلَّمَ بِالْقَلَمِ ۙ﴿٤﴾ عَلَّمَ الْاِنْسَانَ
مَا لَمْ يَعْلَمْ ؕ﴿٥﴾

Artinya :

"Bacalah dengan (menyebut) nama Tuhanmu yang menciptakan!. Dia menciptakan manusia dari segumpal darah. Bacalah! Tuhanmulah Yang Maha mulia. Yang mengajar (manusia) dengan pena. Dia mengajarkan manusia apa yang tidak diketahuinya."

Agar proses pembelajaran berjalan dengan baik dengan kesiapan dan kecepatan dalam umpan balik ke siswa maka pada proses penilaian pekerjaan dibutuhkan suatu sistem yang dapat menilai jawaban siswa yang berupa jawaban esai dengan cara otomatis tetapi tetap mempertahankan keaakuratan, kekonsistenan dan keefisienan serta adil, hal ini sesuai dengan surat An Nisa ayat 58:


۞ إِنَّ ٱللَّهَ يَأۡمُرُكُمۡ أَن تُؤَدُّواْ ٱلۡأَمَٰنَٰتِ إِلَىٰٓ أَهۡلِهَا وَإِذَا
حَكَمۡتُم بَيۡنَ ٱلنَّاسِ أَن تَحۡكُمُواْ بِٱلۡعَدۡلِۚ إِنَّ ٱللَّهَ نِعِمَّا يَعِظُكُم بِهِۦٓۗ إِنَّ
ٱللَّهَ كَانَ سَمِيعَۢا بَصِيرٗا ٥٨

Artinya :

"Sesungguhnya Allah menyuruh kamu menyampaikan amanat kepada yang berhak menerimanya, dan (menyuruh kamu) apabila menetapkan hukum di antara manusia supaya kamu menetapkan dengan adil. Sesungguhnya Allah memberi pengajaran yang sebaik-baiknya kepadamu. Sesungguhnya Allah adalah Maha Mendengar lagi Maha Melihat"

Ayat diatas menjelaskan tentang pentingnya amanah dan keadilan dalam menetapkan suatu keputusan. Ayat ini memiliki kaitan dengan sistem yang

dikembangkan bertujuan membantu proses penilaian agar lebih objektif, konsisten, dan adil bagi setiap peserta didik. Selain itu, penggunaan teknologi dalam penilaian juga harus dilakukan secara bertanggung jawab agar hasil yang diberikan sesuai dengan kemampuan siswa yang sebenarnya.

Penelitian juga sejalan dengan hadis Rasulullah yaitu Rasulullah Shallallahu ‘alaihi wa sallam bersabda:

طَلَبُ الْعِلْمِ فَرِيضَةٌ عَلَى كُلِّ مُسْلِمٍ

“Menuntut ilmu itu wajib bagi setiap muslim” (HR. Ibnu Majah, No. 224).

Hadis tersebut menunjukkan bahwa pengembangan teknologi dalam bidang pendidikan merupakan bagian dari penerapan ilmu untuk kemaslahatan manusia. Selain itu, konsep amanah dalam hadis juga berkaitan dengan tanggung jawab peneliti dalam mengembangkan sistem penilaian yang objektif, akurat, dan adil bagi peserta didik.

Disinilah dibutuhkan teknologi pemrosesan bahasa alami atau sering disebut Natural Language Processing dan machine learning menawarkan dengan membangun sistem prediksi nilai untuk jawaban esai yang dapat membantu guru dalam menghasilkan skor yang lebih seragam dan memberikan umpan balik lebih cepat. Pada awalnya prediksi nilai secara otomatis umumnya menggunakan teknik representasi teks sederhana yang hanya menghitung frekuensi tanpa melihat urutan maupun konteks seperti bag-of-word atau n-gram sehingga sulit menangkap makna kalimat dan variasi bahasa (Hartoko & Leong, 2023). Kemudian munculah word embeddings yang dapat digunakan dalam perhitungan kesamaan semantic dari sebuah kata dan fasttext dapat mengekspresikan kata

sebagai subword atau gabungan dari potongan karakter sehingga dapat mengenali kemiripan dari bentuk kata yang pada akhirnya dapat membantu dalam mengklasifikasi teks diluar Pustaka (Sani & Sarwani, 2022)(Khomsah et al., 2022). Sedangkan long short-term memory (LSTM) merupakan jaringan saraf tiruan yang dirancang untuk memproses urutan data dan mampu mengingat konteks jangka panjang sehingga dapat mengatasi masalah *vanishing gradient* (Malashin et al., 2024; Waqas & Humphries, 2024). Dalam prediksi nilai jawaban esai ini konteks dan kesinambungan kalimat menjadi unsur penting, dengan demikian menggabungkan fasttext dengan LSTM diharapkan mampu mengurangi subjektivitas penilaian dan menghasillkan skor yang konsisten pada jawaban yang serupa.

Terdapat penelitian dengan tujuan yang sama yaitu (Pradani & Suadaa, 2023). peneliti menggunakan dataset sebanyak 532 jawaban esai dari 76 mahasiswa dengan model indoBERT dan terbukti unggul dalam menilai jawaban esai. Peneliti lain (Mizumoto & Eguchi, 2023) menggunakan metode LLM dalam penelitiannya dan mampu menangkap konteks semantik esai lebih mendalam. Selain itu ada peneliti lain (Rajagede, 2021) yang menggunakan data berupa jawaban esai siswa (Ukara dataset) untuk mengklasifikasi jawaban salah atau benar. Peneliti menggunakan metode SBERT dalam penyelesaiannya dan didapatkan hasil lebih akurat, lebih ringan, dan lebih mudah di-deploy dibanding model kompleks sebelumnya. Lalu ada juga penelitian lain yang dilakukan oleh (Buditjahjanto et al., 2022) yang menggunakan Backpropagation Neural Network (BPNN) untuk memprediksi nilai dengan rentang nilai 0-1. Hasil dari penelitian

ini didapatkan tingkat akurasi sebesar 90.63% hal ini membuktikan sistem terbukti efektif, akurat, efisien, dan lebih objektif.

Urgensi penelitian ini terletak pada kebutuhan mendesak akan sistem penilaian jawaban esai yang otomatis, akurat, konsisten, dan efisien. Penelitian penerapan LSTM untuk mendapatkan penilaian skor jawaban soal esai perlu dilakukan secara otomatis karena saat ini proses pemberian skor masih banyak dilakukan secara manual, sehingga membutuhkan waktu lama dan berpotensi kurang konsisten (Hussein et al., 2019; Ramesh & Sanampudi, 2022). Setiap guru mesti harus melakukan penilaian rubrik dengan menggunakan excel, sehingga perlu untuk waktu yang lebih bagi guru untuk melakukan pekerjaan ini, guru masih harus menilai jawaban siswa secara manual menggunakan rubrik di Excel. Proses ini membutuhkan waktu lebih lama karena penilaian dengan rubrik memerlukan pemeriksaan kriteria dan pemberian skor secara teliti (Orrell, 2018). Dengan penerapan LSTM pada skor jawaban esai akan mempersingkat waktu, kedua guru pada waktu melakukan penilaian harus akurat tetapi dengan kondisi skorsing menggunakan excel dan melihat nilai akurasi tidak lah tepat, maka dengan menggunakan model LSTM berharap mendapatkan akurasi yang dipertanggung jawabkan. Penelitian ini penerapan model LSTM pada penilaian jawaban esai dapat membantu guru menilai dengan lebih cepat, konsisten, dan akurat. Penilaian manual menggunakan Excel masih berisiko kurang objektif karena guru harus membaca dan memberi skor satu per satu. Dengan LSTM, sistem dapat mempelajari pola urutan kata dalam jawaban esai (Sherstinsky, 2020) sehingga hasil penilaian diharapkan lebih terukur dan dapat

dipertanggungjawabkan (Hochreiter & Schmidhuber, 1997; Ramesh & Sanampudi, 2022; Faseeh et al., 2024).

Oleh karena itu, pengembangan model prediksi nilai jawaban esai yang menggabungkan FastText dan LSTM menjadi penting dan relevan untuk menghasilkan penilaian yang lebih objektif, konsisten, dan cepat, sekaligus menjawab kebutuhan praktis guru dan dosen dalam mengevaluasi jawaban esai secara efisien.

1.2 Pernyataan Masalah

Pernyataan masalah pada penelitian ini adalah sebagai berikut:

1. Bagaimana cara mengimplementasikan fasttext dan algoritma LSTM dalam memprediksi skor jawaban esai otomatis, akurat, konsisten, dan efisien?
2. Bagaimana cara memprediksi soal esay otomatis, akurat, konsisten, dan efisien menggunakan fasttext dan algoritma LSTM?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Untuk mengimplementasikan fasttext dan algoritma LSTM dalam memprediksi skor jawaban esai otomatis, akurat, konsisten, dan efisien.
2. Memprediksi soal esay otomatis, akurat, konsisten, dan efisien menggunakan fasttext dan algoritma LSTM.

1.4 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah sebagai berikut:

a. Membantu guru atau dosen dalam memberi skor pada jawaban esai dan pengefesienan dalam penggunaan waktu.
b. Guru atau dosen dapat memberikan penilaian dengan cepat meskipun dalam jumlah besar.
c. Siswa atau mahasiswa mendapat umpat balik lebih cepat.

1.5 Batasan Masalah

a. Data yang digunakan pada penelitian ini berupa teks dari jawaban esai siswa SMP kelas 9 pada mata pelajaran Bahasa Inggris.
b. Penelitian ini berfokus pada pembuatan model menggunakan metode *fasttext* dan *LSTM*.

BAB II

STUDI PUSTAKA

2.1 Prediksi Nilai Jawaban Esai

Penelitian yang dilakukan oleh (Pradani & Suadaa, 2023) tentang koreksi jawaban esai secara otomatis. Penelitian ini timbul karena dalam memberikan skor pada jawaban esai secara manual membutuhkan waktu, tenaga dan sering tidak konsisten sehingga dapat menimbulkan bias dalam penilaian. Untuk itu peneliti menggunakan koreksi jawaban esai otomatis dengan menggunakan pendekatan semantic textual similarity berbasis Transformer indoBERT dengan menggunakan data jawaban esai dari mahasiswa dan jawaban dari dosen yang nantinya output dari penelitian ini akan dibandingkan dengan jawaban dosen dan direpresentasikan dalam rentang 1-10. Dataset: 532 esai dari 76 mahasiswa, 7 soal. Untuk pra pemrosesan data, peneliti menggunakan case folding, tokenizing, stop removal , dan stemming. Dari hasil tersebut akan di proses lebih lanjut menggunakan 3 perlakuan, yang pertama menggunakan TF-IDF dengan cosine similarity kemudian yang kedua dengan TF-IDF dengan Linear Regression dan yang terakhir menggunakan fine tuned indoBERT (transformer) untuk menghitung semantic similarity dan prediksi skor. Untuk tahap evaluasi menggunakan MAE (Mean Absolut Error) dan RMSE (Root Mean Square Error) serta 5-fold cross validation. Hasil yang didapatkan dalam penelitian ini untuk TF-IDF dengan cosine similarity menghasilkan nilai MAE sebesar 0,437 dan RMSE sebesar 0,525. Sedangkan untuk TD-IDF dengan linear regression didapatkan nilai

MAE 0.208 dan RMSE 0.295 untuk penggunaan indoBERT (fine-tuned) didapatkan MAE sebesar 0.128 dan RMSE sebesar 0.2. sehingga model IndoBERT (fine-tuned) menghasilkan nilai terbaik dan terbukti unggul dalam menilai jawaban esai karena mampu memahami teks, bukan hanya sekedar kesamaan kata.

Peneliti lain oleh (Mizumoto & Eguchi, 2023) juga meneliti tentang koreksi jawaban otomatis menggunakan LLM. Penelitian ini muncul karena penilaian esai manual sering tidak konsisten, memakan waktu dan biaya tinggi serta terdapat sistem Automated Essay Scoring yang masih berbasis *hand-crafted features* dan kurang fleksibel. Dataset pada peneitian ini berupa data esai dari dataset ASAP (Automated Student Assessment Prize) dan representasi teks yang diproses oleh pre-trained AI language models yang nantinya hasil akan diprediksi dan dibandingkan dengan skor manual. Sebelum data diproses data harus di bersihkan terlebih dahulu dan tokenisasi sesuai model bahasa (LLM) dan normalisasi panjang input agar sesuai dengan kapasitas model. Dari hasil ini akan dilakukan AI Language Models (LLM) termasuk GPT dan transformer base models. Setelah itu fine tuning pada dataset AES (*automated essay scoring*) dan untuk evaluasi menggunakan Quadratic Weighted Kappa (QWK) dan juga korelasi. Hasil yang didapat pada penelitian ini model berbasis AI Language Model (GPT) menghasilkan performa lebih baik. LLM mampu menangkap konteks semantik esai lebih mendalam.

Selain itu peneliti lain yang berjudul *improving automatic essay scoring for indonesian language using simpler model dan richer feature* oleh (Rajagede,

2021) yang menggunakan data berupa jawaban esai siswa (Ukara dataset) untuk mengklasifikasi jawaban salah atau benar. Dalam tahap pra-pemrosesan, penulis menerapkan menghapus kata sering muncul ("yang", "lebih", "akan", dll) dan SMOTE (Synthetic Minority Oversampling) untuk menyeimbangkan kelas minoritas (hanya pada Question B), selain itu juga membagi dataset menjadi klas train, validation, test (sesuai Ukara 1.0 Challenge). Metode utama yang digunakan adalah SBERT menghasilkan embedding 768-dimensi dan Neural network single hidden layer dilatih menggunakan hyperparameter tuning (Optuna TPE). Setelah itu dilakukan evaluasi berupa prediksi skor (true/false) dengan F1-score pada test set. Hasil penelitian menunjukkan bahwa F1-Score rata-rata 0.829 (lebih tinggi dari fastText + stacking: 0.821). model sederhana dengan fitur kaya (SBERT) lebih akurat, lebih ringan, dan lebih mudah di-deploydibanding model kompleks sebelumnya

Penelitian yang dilakukan oleh (Buditjahjanto et al., 2022) bertujuan membangun sistem AES berbasis Backpropagation Neural Network (BPNN) untuk memprediksi nilai dengan rentang nilai 0-1 dan mengklasifikasikan tingkat kompetensi peserta ujian dalam tingkat komptensi rendah, sedang, dan tinggi. Dataset yang digunakan berupa jawaban esai peserta ujian sebanyak 71 data dari 12 pertanyaan dalam 3 unit kompetensi. Sebelum data digunakan peneliti menggunakan *case folding, tokenizing, stopword, dan stemming* bahasa Indonesia (library sastrawi) lalu dilakukan pembobotan teks dengan TF-IDF, kemudian penggunaan Backpropagation Neural Network (BPNN) dengan arsitektur prediksi skor 12-4-1 dan klasifikasi level kompetensi 3-

4-1. Untuk evaluasi prediksi skor menggunakan *Mean Absolute Error* (MAE) dan *Accuracy*, sedangkan untuk klasifikasi kompetensi menggunakan Confusion Matrix (Accuracy, Precision, Recall, F1-Score). Hasil penelitian menunjukkan bahwa untuk prediksi skor kompetensi nilai MAE sebesar 0.061621 dan accuracy sebesar 97.97%. dan untuk klasifikasi kompetensi didapatkan nilai accuracy sebesar 90.63%, Precision sebesar 91.67%, Recall sebesar 88.88%, dan F1-Score sebesar 88.57%. hal ini membuktikan sistem terbukti efektif, akurat, efisien, dan lebih objektif.

Penelitian oleh (Chamidah et al., 2022) menggunakan data berupa jawaban esai uji dan kunci jawaban (*essay test answer & reference answer*) dalam bentuk teks, peneliti menerapkan *case folding* kemudian menvektorkan kata menggunakan pretrained SBERT. Kemudian peneliti menggunakan model SBERT (Siamese BERT) untuk menghasilkan representasi semantik dari teks esai dan kunci jawaban, sebagai vektor panjang 512. Lalu menghitung Cosine Similarity antara embedding kunci jawaban dan jawaban uji untuk mendapatkan nilai similaritas semantik yang kemudian dijadikan output model (nilai esai otomatis). Setelah itu dilakukan tahap pasca pemrosesan berupa output similarity dibandingkan dengan nilai yang diperoleh manual, kemudian dihitung MAE (*Mean Absolute Error*) dan Pearson Correlation sebagai evaluasi. Hasil penelitian menunjukkan bahwa penelitian menghasilkan rata-rata MAE sebesar 0,26 dan rata-rata korelasi Pearson sebesar 0,78 antara nilai otomatis dan nilai evaluator manusia. Model cukup efektif, tetapi masih perlu peningkatan dengan dataset lebih besar atau fine-tuning SBERT.

Penelitian yang dilakukan oleh (Kumar & Boulanger, 2020) menggunakan data berupa Dataset *Automated Student Assessment Prize* (ASAP) – *Grade 7 essays* (narrative, 1567 essays) + 1592 *linguistic indices (cohesion, lexical diversity, syntactic sophistication,* dsb.), dengan output yang ditargetkan yaitu *Rubric scores (Ideas, Organization, Style, Conventions*) dan *Holistic score* (0–24). Dalam tahap pra-pemrosesan, penulis menerapkan *feature selection* yaitu dengan melakukan reduksi dari 1592 fitur menjadi 282 fitur, normalisasi, pruning fitur rendah variansi, regularisasi Lasso & Ridge (ElasticNet). Metode utama yang digunakan adalah Deep Learning (Multi-Layer Perceptron Neural Network dengan 2–6 hidden layers, ensemble models) + *Explainable* AI (SHAP: KernelSHAP, DeepSHAP, GradientSHAP). Setelah itu dilakukan tahap evaluasi menggunakan akurasi prediksi dan trustworthiness explanation models (*precision, recall*, F1, QWK). Hasil penelitian menunjukkan bahwa Quadratic Weighted Kappa (QWK) memberikan nilai Holistic sebesar 0.785; Rubrics: Ideas sebesar 0.731, Organization sebesar 0.676, Style sebesar 0.650, Conventions sebesar 0.674, dan Descriptive accuracy meningkat ±10% dengan penambahan hidden layers, SHAP (Deep/Grad) jauh lebih cepat dibanding KernelSHAP.

Penelitian oleh (Ramesh & Sanampudi, 2022) yang juga meneliti tentang koreksi jawaban esai otomatis dengan menggunakan data berupa berbagai dataset esai (e.g., ASAP, TOEFL11), fitur linguistik & semantik yang dipakai di setiap studi dan output yang ditargetkan dalam penelitian ini yaitu skor esai otomatis (prediksi mendekati skor manusia). Dalam tahap pra-pemrosesan, penulis menerapkan identifikasi langkah umum yaitu tokenisasi*, stemming/lemmatization*,

stopword removal, dan vectorization (TF-IDF/word *embeddings*), sedangkan metode utama yang digunakan adalah ML klasik (SVM, RF), DL (CNN, LSTM, BERT), *feature engineering & hybrid approach*. Setelah itu dilakukan tahap pasca-pemrosesan berupa Analisis evaluasi model, perbandingan hasil dengan human rater, identifikasi gap riset. Hasil penelitian menunjukkan bahwa review menyimpulkan model DL (BERT-based) unggul tetapi masalah explainability & bias masih menjadi tantangan.

Penelitian lain oleh (Sani & Sarwani, 2022) yang juga meneliti tentang koreksi jawaban esai otomatis menggunakan data berupa jawaban esai sekolah menengah pertama mata pelajaran bahasa Indonesia. Pada tahap pra pemrosesan data dilakukan *case folding* lalu tokenizing kemudian perluas korpus dengan wikipedia dengan dimensi 300 dan jendela 5 setelah itu dilakukan normalisasi vektor. Metode utama pada penelitian ini adalah jaringan saraf backpropagation dengan arsitektur 300-128-32-8-1, kemudian aktivasi ReLU dan optimasi SGD dengan scema training atau testing beberapa kombinasi. Untuk evaluasi digunakan nilai MAPE terkecil serta analisis potensi overfitting saat porsi data latih kecil. Hasil yang didapat pada penelitian ini adalah ombinasi terbaik pada saat 75% training (375 data) / 25% testing (125 data). Untuk kinerja MAPE training sebesar 5% dan MAPE testing sebesar 8% seingga model dinilai “sangat baik” untuk koreksi esai otomatis.

Penelitian (Misgna et al., 2025) yang mereview penggunaan metode deep learning dalam koreksi jawaban otomatis yang kemudian disintesis temuannya berupa arsitektur, representasi esai, metrik evaluasi serta celah riset. Dalam

penelitian ini untuk representasi esai didapatkan handcrafted features, one-hot, embeddings seperti Word2Vec/GloVe/FastText, dan POS. Sementara untuk kategori model didapat CNN/RNN/LSTM, attention, Transformer/BERT. Hasil yang didapat pada penelitian ini Model DL mampu memprediksi skor esai secara end-to-end, namun keterjelasan atau penjelasan dan feedback masih terbatas; literatur feedback generation relatif sedikit (data terbatas). BERT/Transformer memberikan kinerja yang kuat, peningkatan lebih lanjut melalui loss gabungan (regresi+ranking) dan pairwise contrastive, atau multi-scale features, tapi terdapat batasan pada panjang token sebesar 512.

Terdapat penelitian lain oleh (Balaha & Saafan, 2021) penilaian otomatis sejalan dengan Automatic Exam Correction Framework (HMB-AECF) dengan Algoritma Similarity Checker. Data yang digunakan berupa jawaban mahasiswa. penelitian ini bertujuan untuk mengevaluasi kesamaan semantik jawaban teks (benar/salah atau tingkat kemiripan). Dalam tahap preprocessing digunakan beberapa metode word2vec, FastText, Glove, Universal Sentence Encoder (USE). Kemudian peneliti menggunakan Automatic Exam Correction Framework (HMB-AECF) yang terdiri dari lima layer untuk menangani proses penilaian otomatis dan algoritma similarity checker untuk teks sedangkan untuk persamaan matematis menggunakan HMB-MMS-EMA. Untuk evaluasi menggunakan akurasi untuk klasifikasi dan Root Mean Square Error (RMSE) untuk penilaian kesalahan / regresi. Hasil yang didapatkan untuk teks akurasi tertinggi sebesar 77,95% dengan menggunakan USE tanpa fine tuning dan RMSE terendah 1,09 dengan menggunakan USE tanpa fine tuning sedangkan untuk persamaan

matematis menggunakan algoritma HMB-MMS-EMA sebesar 100% dibandingkan dengan SymPy hanya mencapai 71,33%.

Berdasarkan analisis terhadap penelitian-penelitian terdahulu, dapat diidentifikasi beberapa peluang penelitian atau research gap yang menjadi dasar penting bagi penelitian ini. Peluang pada penelitian ini adalah mencari pemilihan metode prediksi yang digunakan dalam penilaian jawaban esai otomatis, khususnya untuk Bahasa Indonesia. Pada penelitian sebelumnya masih banyak berfokus pada penggunaan model prediksi berbasis Transformer, seperti penggunaan metode IndoBERT dan SBERT, karena model tersebut memiliki kemampuan yang baik dalam memahami konteks bahasa (Koto et al., 2020). Selain itu, beberapa penelitian juga masih menggunakan jaringan saraf sederhana, seperti Backpropagation Neural Network. Pemanfaatan FastText yang dikombinasikan dengan *Long Short-Term Memory* (LSTM) dalam penilaian jawaban esai Bahasa Indonesia masih relatif terbatas karena sebagian besar penelitian sebelumnya lebih banyak menggunakan metode lain, seperti BERT/IndoBERT, similarity, atau FastText dengan algoritma selain LSTM. Padahal, FastText mampu menangkap informasi subword, sedangkan LSTM dapat mempelajari urutan kata dalam teks esai. Oleh karena itu, kombinasi FastText dan LSTM masih memiliki peluang untuk dikembangkan dalam penelitian penilaian esai otomatis Bahasa Indonesia (Rajagede, 2021; Pradani et al., 2023; Aliyah et al., 2025; Bojanowski et al., 2017).

Berdasarkan peluang penelitian tersebut, maka penelitian yang dikerjakan adalah untuk mengembangkan model prediksi nilai jawaban esai berbasis

kombinasi FastText dan LSTM dengan memanfaatkan gabungan teks soal dan jawaban sebagai input. FastText digunakan untuk menghasilkan representasi kata yang lebih fleksibel terhadap variasi bahasa, kesalahan ejaan, dan kata di luar kosakata, sedangkan LSTM digunakan untuk mempelajari urutan kata serta hubungan antarbagian dalam teks. Dengan pendekatan ini, penelitian diharapkan dapat memberikan kontribusi dalam pengembangan sistem penilaian jawaban esai otomatis yang lebih akurat, efisien, konsisten, dan sesuai dengan karakteristik Bahasa Indonesia.

Keterbaruan pada penelitian ini adalah penggunaan input yang berupa gabungan soal dan jawaban. Selain itu, pemodelan teks dan metode utama yang akan digunakan adalah fasttext dengan kombinasi *Long Short-Term Memory* (LSTM) dimana belum diteliti pada penelitian sebelumnya. Referensi atau jurnal yang digunakan pada penelitian ini secara rinci dapat dilihat pada tabel 2.1

Tabel 2. 1 Daftar Jurnal

No.	Sumber	Variabel bebas (input)	Metode	Variabel terikat (output)	Hasil
1.	Pradani & Suadaa (2023)	Data: 532 esai dari 76 mahasiswa, 7 soal; pasangan jawaban mahasiswa & jawaban acuan dosen.	Pendekatan utama: Semantic Textual Similarity (STS). Model: baseline TF-IDF (cosine & linear regression) vs IndoBERT fine-tuned untuk menghitung kesamaan semantik & memprediksi skor. Evaluasi: 5-fold cross-validation; metrik MAE & RMSE.	Skor otomatis jawaban esai pada rentang 1–10 (diprediksi model) yang dibandingkan dengan skor dosen (ground truth).	TF-IDF + Cosine: MAE 0,437; RMSE 0,525. TF-IDF + Linear Regression: MAE 0,208; RMSE 0,295. IndoBERT (fine-tuned): MAE 0,128; RMSE 0,200 terbaik. Kesimpulan: IndoBERT unggul karena mampu memahami makna/semantik, bukan sekadar kemiripan kata.
2.	Mizumoto & Eguchi (2023)	Dataset: esai dari ASAP (Automated Student Assessment Prize). Pra-proses: pembersihan teks, tokenisasi sesuai LLM, normalisasi panjang input. Variabel bebas/perlakuan: jenis model (LLM berbasis GPT vs transformer base models lain), representasi teks dari model pra-latih, serta fine-tuning pada	Alur: (1) pra-pemrosesan, (2) ekstraksi representasi menggunakan pre-trained AI language models (LLM, termasuk GPT), (3) fine-tuning untuk tugas automated essay scoring (AES), (4) evaluasi menggunakan Quadratic Weighted Kappa (QWK) dan korelasi terhadap skor manusia.	Skor esai otomatis (prediksi model) yang dibandingkan dengan skor manual/penilai manusia	Model LLM berbasis GPT memberikan performa lebih baik dibanding pendekatan lain; LLM mampu menangkap konteks semantik esai lebih mendalam. (Teks tidak menyertakan angka QWK/korelasi spesifik.)

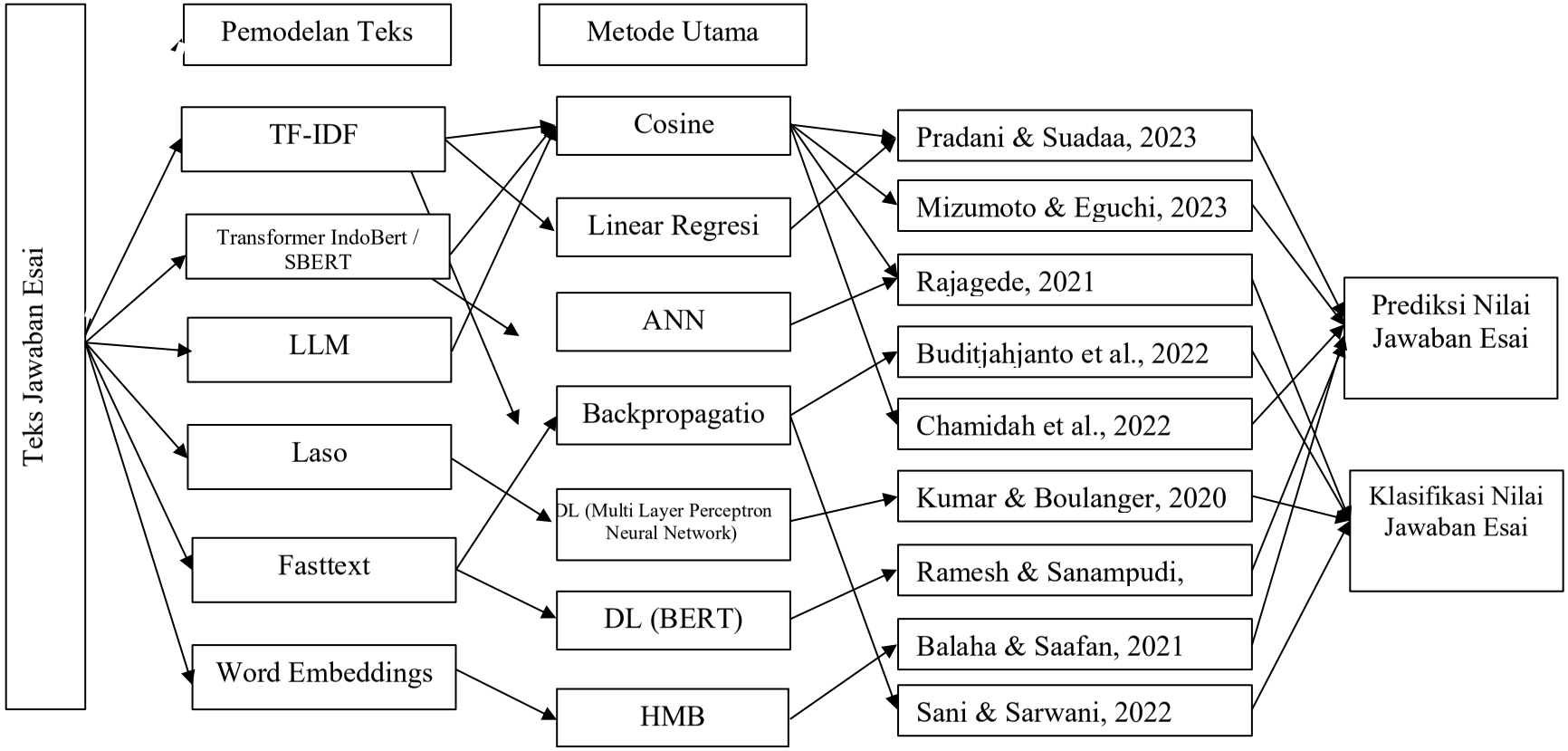
No.	Sumber	Variabel bebas (input)	Metode	Variabel terikat (output)	Hasil
		dataset AES.			
3.	Rajagede (2021)	Dataset: Ukara (jawaban esai siswa) untuk klasifikasi benar/salah. Pra-proses: hapus kata frekuen (“yang”, “lebih”, “akan”, dll), SMOTE (khusus Question B) untuk menyeimbangkan kelas, split train/val/test sesuai Ukara 1.0 Challenge. Perlakuan/variabel bebas: representasi SBERT (embedding 768-dimensi); pembandingan dengan baseline fastText + stacking. Tuning: Optuna (TPE).	Arsitektur: Neural network 1 hidden layer di atas embedding SBERT; hyperparameter tuning (Optuna TPE).	Label prediksi: klasifikasi benar /salah (true/false) jawaban esai (AES).	F1-score rata-rata = 0,829, lebih tinggi daripada fastText + stacking = 0,821. Kesimpulan: model sederhana + fitur kaya (SBERT) lebih akurat, ringan, dan mudah di-deploy dibanding model kompleks sebelumnya.
4.	Buditjahjanto et al., 2022	Jawaban esai peserta ujian (71 data) dari 12 pertanyaan dalam 3 unit kompetensi: - U1: Occupational	Untuk preprocessing digunakan metode antara lain : Case Folding (huruf kecil), Tokenizing (memecah teks menjadi kata, Stopword Removal, Stemming Bahasa	Prediksi skor kompetensi (nilai 0–1) dan Klasifikasi tingkat	Prediksi skor kompetensi: MAE = 0.061621, Accuracy = 97.97% - Klasifikasi kompetensi: Accuracy = 90.63%, Precision = 91.67%, Recall = 88.88%, F1-

No.	Sumber	Variabel bebas (input)	Metode	Variabel terikat (output)	Hasil
		Safety (X1–X4) - U2: Software & Hardware (X5–X8) - U3: Digital Image & Multimedia (X9–X12)	Indonesia (library Sastrawi) Untuk metode utama menggunakan - Pembobotan teks dengan TF-IDF Sedangkan evaluasi performa dengan menggunakan : - Prediksi skor: Mean Absolute Error (MAE), Accuracy - Klasifikasi kompetensi: Confusion Matrix (Accuracy, Precision, Recall, F1-Score)	kompetensi: rendah, sedang, tinggi	Score = 88.57% - Sistem terbukti efektif, akurat, efisien, dan lebih objektif dibanding penilaian manual
	Chamidah et al., 2022	Jawaban esai uji dan kunci jawaban (essay test answer & reference answer) dalam bentuk teks	Untuk preprocessing menggunakan Case folding (mengubah teks ke huruf kecil). Transformasi teks menjadi embedding/vektor menggunakan SBERT (pretrained) setelah praproses teks. Untuk metode utama menggunakan Model SBERT (Siamese BERT). Menghitung Cosine Similarity antara embedding kunci jawaban dan jawaban uji untuk mendapatkan nilai similaritas semantik yang kemudian dijadikan output model (nilai esai otomatis	Nilai esai otomatis yang dihasilkan oleh model, dibandingkan dengan nilai evaluator manusia; serta ukuran performa yaitu MAE (Mean Absolute Error) dan Pearson Correlation	Hasil yang dicapai: • Rata-rata MAE sebesar **0,26**. • Rata-rata korelasi Pearson sekitar **0,78**. • Dengan demikian, sistem berbasis SBERT menunjukkan potensi yang baik untuk membantu atau mempercepat penilaian esai pendek, meningkatkan konsistensi (karena otomatis) dan efisiensi (lebih cepat daripada manual) seperti yang menjadi latar belakang penelitian
5.	Kumar & Boulanger, 2020	Dataset Automated Student Assessment Prize (ASAP) – Grade 7 essays (narrative,	Untuk preprocessing digunakan Feature selection (reduksi 1592 → 282 fitur), normalisasi, pruning fitur rendah variansi, regularisasi Lasso &	Rubric scores (Ideas, Organization, Style,	Quadratic Weighted Kappa (QWK): Holistic = 0.785; Rubrics: Ideas 0.731, Organization 0.676, Style 0.650,

No.	Sumber	Variabel bebas (input)	Metode	Variabel terikat (output)	Hasil
		1567 essays) + 1592 linguistic indices (cohesion, lexical diversity, syntactic sophistication, dsb.	Ridge (ElasticNet). Untuk media utama menggunakan Deep Learning (Multi-Layer Perceptron Neural Network, 2–6 hidden layers, ensemble models) + Explainable AI (SHAP: KernelSHAP, DeepSHAP, GradientSHAP). Untuk evaluasi akurasi prediksi & trustworthiness explanation models (precision, recall, F1, QWK)	Conventions) dan Holistic score (0–24)	Conventions 0.674, Descriptive accuracy meningkat ±10% dengan penambahan hidden layers, SHAP (Deep/Grad) jauh lebih cepat dibanding KernelSHAP tanpa mengorbankan akurasi buat kalimat saja jangan dibuat point
6.	Ramesh & Sanampudi, 2022	Berbagai dataset esai (e.g., ASAP, TOEFL11), fitur linguistik & semantik yang dipakai di setiap studi	Identifikasi langkah umum preprocessing: tokenisasi, stemming/lemmatization, stopword removal, vectorization (TF-IDF/word embeddings), Ringkasan metode utama: ML klasik (SVM, RF), DL (CNN, LSTM, BERT), feature engineering & hybrid approach,	Skor esai otomatis (prediksi mendekati skor manusia)	Review menyimpulkan model DL (BERT-based) unggul tetapi masalah explainability & bias masih menjadi tantangan; memberi roadmap penelitian ke depan
7.	Sani & Sarwani, 2022	kalimat diperoleh dengan FastText (embedding) lalu dirata-rata dari vektor kata.	Untuk tahap preprocessing dilakukan case folding , tokenizing, perluasan korpus dengan Wikipedia; lalu normalisasi vektor (rata-rata vektor kata → vektor kalimat). Untuk metode utama menggunakan FastText (arsitektur CBOW) untuk embedding; model pra-latih Wikipedia (dimensi 300; jendela 5) ditambah pelatihan pada data penelitian.Jaringan saraf	Skor prediksi jawaban (Y_pred) yang dibandingkan dengan skor guru (Y_aktual)	Kombinasi terbaik: 75% training (375 data) / 25% testing (125 data). Kinerja: MAPE training = 5% dan MAPE testing = 8% → model dinilai “sangat baik” untuk koreksi esai otomatis

No.	Sumber	Variabel bebas (input)	Metode	Variabel terikat (output)	Hasil
			backpropagation (arsitektur 300–128–32–8–1, aktivasi ReLU, optimisasi SGD), dengan skema training/testing beberapa kombinasi. Sedangkan untuk pemilihan kombinasi train/test terbaik berdasar MAPE terkecil serta analisis potensi overfitting saat porsi data latih kecil		
8.	Balaha & Saafan, 2021	Jawaban mahasiswa (teks) dan persamaan matematis; variasi representasi (Word2Vec, FastText, GloVe, USE) sebagai perlakuan/model	Pembersihan teks (implisit), tokenisasi (implisit), pembuatan embedding dengan Word2Vec/FastText/GloVe/USE (tanpa fine-tuning untuk skenario terbaik). Untuk metode utama menggunakan HMB-AECF (5 layer) untuk alur penilaian otomatis; algoritma similarity checker untuk teks; HMB-MMS-EMA untuk persamaan matematis Untuk evaluasi menggunakan Konversi skor kemiripan → label benar/salah atau tingkat kemiripan; pemetaan skor/regresi; pemilihan model terbaik (berdasarkan akurasi/RMSE)	Prediksi kelas (benar/salah) atau tingkat kemiripan (regresi) dibanding ground truth	eks: akurasi tertinggi 77,95% (USE, tanpa fine-tuning); RMSE terendah 1,09 (USE, tanpa fine-tuning). Persamaan matematis: HMB-MMS-EMA = 100%, vs SymPy = 71,33%.

2.2 Kerangka Teori



Gambar 2. 1 Kerangka Teori

2.3 Penilaian

2.3.1 Konsep Dasar Penilaian Hasil Belajar

a. Pengertian penilaian

Berdasarkan literatur terkini, penilaian dalam konteks pendidikan memiliki definisi yang komprehensif. Van der Vleuten dan Schuwirth mendefinisikan penilaian sebagai "any formal or purported action to obtain information about the competence and performance of a student" (Gupta, 2023). Definisi ini menekankan bahwa penilaian merupakan tindakan sistematis untuk mengumpulkan informasi tentang kompetensi dan kinerja peserta didik. (Rao & Banerjee, 2023) memberikan perspektif yang lebih luas dengan menyatakan bahwa "Classroom assessment is the process of documenting the knowledge, skills, attitudes and beliefs of learners" . Definisi ini menekankan fungsi dokumentasi yang komprehensif dalam proses pembelajaran, tidak hanya terbatas pada aspek kognitif tetapi juga mencakup dimensi afektif dan psikomotor. Dalam konteks pendidikan Indonesia, (Munaroh, 2024) menjelaskan bahwa asesmen atau penilaian merupakan "proses pengumpulan dan pengolahan informasi untuk mengukur pencapaian hasil belajar peserta didik yang mencakup penilaian otentik, penilaian diri, penilaian berbasis portofolio, ulangan, ulangan harian, ulangan tengah semester, ulangan akhir semester, ujian tingkat kompetensi, ujian mutu tingkat kompetensi, ujian nasional, dan ujian sekolah/madrasah".

Tujuan Evaluatif dan Monitoring Penilaian digunakan untuk memonitor kemajuan belajar peserta didik, memberikan umpan balik untuk perbaikan, dan mendukung keputusan kurikulum serta kebijakan pendidikan, termasuk promosi atau seleksi (Gupta, 2023). Tujuan Informatif untuk Pengajaran (Rao & Banerjee, 2023) menekankan bahwa penilaian kelas berfungsi untuk "documenting the knowledge, skills, attitudes and beliefs of learners" yang memberikan informasi penting bagi guru untuk memperbaiki strategi pengajaran dan pembelajaran (Munaroh, 2024). Tujuan Pengembangan Kompetensi Dalam konteks pendidikan berbasis kompetensi, penilaian bertujuan untuk mengukur pencapaian kompetensi yang telah ditetapkan dalam kurikulum dan memberikan gambaran holistik tentang kemampuan peserta didik.

Penilaian memiliki beberapa fungsi kunci dalam meningkatkan kualitas pembelajaran yaitu Fungsi Formatif dan Sumatif Penilaian formatif memberikan umpan balik berkelanjutan untuk mengarahkan proses pengajaran, sedangkan penilaian sumatif menilai pencapaian hasil belajar dan memberikan masukan pada desain kurikulum (Rao & Banerjee, 2023). (Putri & Zakir, 2023) menegaskan bahwa penilaian formatif adalah "kegiatan penilaian berulang yang menghasilkan umpan balik untuk memperbaiki proses belajar dan pengajaran". Lalu fungsi Peningkatan Akuntabilitas Pengajaran Literatur menunjukkan bahwa penilaian yang terencana dengan baik meningkatkan tanggung

jawab pengajaran dan mendorong perbaikan praktik pengajaran secara berkelanjutan (Gupta, 2023). Kemudian ada fungsi Diagnostik dan Remedial Penilaian berfungsi sebagai alat diagnostik untuk mengidentifikasi kesulitan belajar peserta didik sehingga intervensi pembelajaran yang tepat dapat dirancang dan diimplementasikan (Munaroh, 2024).

b. Jenis-jenis penilaian dalam pembelajaran

Jenis-jenis penilaian ada 4 kategori yaitu Penilaian Formatif, Sumatif, Diagnostik, dan Penilaian Autentik**.**

- Penilaian Formatif (Putri & Zakir, 2023) mendefinisikan penilaian formatif sebagai "kegiatan penilaian berulang yang menghasilkan umpan balik untuk memperbaiki proses belajar dan pengajaran". Penilaian ini dilakukan secara berkelanjutan selama proses pembelajaran berlangsung dengan tujuan memberikan informasi untuk perbaikan pembelajaran. (Gu, 2021) memperkuat konsep ini dengan menyatakan bahwa penilaian formatif memerlukan kerangka validasi berbasis argumen untuk memastikan bahwa umpan balik yang diberikan benar-benar efektif dalam meningkatkan pembelajaran.
- Penilaian Sumatif Penilaian sumatif "mengevaluasi pencapaian hasil belajar pada akhir periode dan digunakan untuk penentuan capaian, promosi, atau akreditasi" (Putri & Zakir, 2023). Jenis penilaian ini biasanya dilakukan pada akhir unit pembelajaran,

semester, atau tahun ajaran untuk mengukur pencapaian kompetensi secara keseluruhan.

- Penilaian Diagnostik (Munaroh, 2024) menjelaskan bahwa penilaian diagnostik berfungsi sebagai "alat untuk mengidentifikasi kesulitan awal peserta didik sehingga intervensi dapat dirancang". Penilaian ini dilakukan sebelum pembelajaran dimulai untuk mengidentifikasi kemampuan prasyarat dan kesulitan belajar yang mungkin dihadapi peserta didik.
- Penilaian Autentik Mattison et al. menyatakan bahwa "Authentic assessment is introduced as the heart of competency-based education" dan menekankan tugas yang mencerminkan situasi dunia nyata yang relevan dengan kompetensi profesional (Anđelković, 2022). Penilaian autentik menggunakan tugas-tugas yang bermakna dan kontekstual yang mencerminkan aplikasi pengetahuan dan keterampilan dalam situasi nyata.

Esai sebagai Alat Penilaian Autentik Penilaian esai memiliki posisi strategis dalam asesmen autentik karena kemampuannya untuk mengukur keterampilan berpikir tingkat tinggi, kemampuan komunikasi tertulis, dan penerapan pengetahuan dalam konteks yang bermakna. (Anđelković, 2022) menunjukkan bahwa penilaian esai akademik melibatkan "*procedure and correspondence*" antara penilaian guru dan peer assessment yang dapat memberikan perspektif holistik tentang kemampuan menulis akademik peserta didik.

Karakteristik Autentik dalam Penilaian Esai Penilaian esai dianggap autentik karena:

- Mencerminkan tugas-tugas yang relevan dengan dunia nyata
- Memungkinkan peserta didik untuk menunjukkan pemahaman mendalam
- Mengintegrasikan berbagai keterampilan (analisis, sintesis, evaluasi) Memberikan kesempatan untuk ekspresi kreatif dan orisinal

(Milian, 2024) dalam studinya tentang peer assessment pada penulisan esai menegaskan bahwa pendekatan ini "menunjukkan potensi pembelajaran tetapi juga menunjukkan bahwa tanpa pelatihan kriteria, validitas dan konsistensi dapat terpengaruh".

c. Prinsip-prinsip penilaian yang baik.

Dalam penilaian terdapat prinsip-prinsip yang harus diperhatikan yaitu Validitas, Reliabilitas, Objektivitas, dan Kepraktisan**.**

- Validitas Gupta (2023) menekankan bahwa "Validity is a unitary concept, and evidence can be drawn from many aspects such as content and construct-related and empirical evidence". Konsep validitas modern tidak lagi dipandang sebagai karakteristik instrumen, melainkan sebagai kualitas interpretasi dan penggunaan skor yang didukung oleh bukti empiris yang komprehensif. (Gu, 2021) mengembangkan kerangka validasi berbasis argumen untuk penilaian formatif, menekankan bahwa validitas bergantung pada

"bukti interpretasi skor dan argumen validitas yang mendukung klaim penafsiran skor". Pendekatan ini mengharuskan pengumpulan bukti dari berbagai sumber untuk mendukung kesimpulan yang ditarik dari hasil penilaian.

- Reliabilitas mencakup konsistensi hasil penilaian. Gupta (2023) menyatakan bahwa reliabilitas merupakan "komponen penting bersama validitas, kepraktisan, dan dampak pendidikan" dalam model utilitas penilaian. Reliabilitas tidak hanya berkaitan dengan konsistensi internal instrumen, tetapi juga konsistensi antarpenilai dan konsistensi temporal.
- Objektivitas dalam penilaian dicapai melalui penggunaan kriteria penilaian yang jelas, rubrik yang terstruktur, dan pelatihan penilai yang memadai. Rao dan Banerjee (2023) menekankan bahwa "kriteria desain instrumen (rubrik jelas, pelatihan penilai) meningkatkan objektivitas dan kepraktisan implementasi dalam konteks kelas" .
- Kepraktisan berkaitan dengan kemudahan implementasi penilaian dalam konteks pembelajaran nyata, termasuk pertimbangan waktu, biaya, dan sumber daya yang tersedia. Penilaian yang praktis adalah yang dapat dilaksanakan secara efisien tanpa mengorbankan kualitas pengukuran.

Dalam melakukan suatu penilaian terdapat Tantangan dalam Menjaga Keadilan dan Konsistensi Penilaian Jawaban Esai

- Variabilitas Penilai (Rater Variability) (Anđelković, 2022) menemukan bahwa "Lower correlation and higher difference between mean grades awarded by teachers and peers ... may indicate the essay writing aspects students are weakest at". Hal ini menunjukkan bahwa tanpa pelatihan dan rubrik yang memadai, perbedaan penilaian antara penilai dapat signifikan, yang berimplikasi pada keadilan dan konsistensi penilaian. (Milian, 2024) memperkuat temuan ini dengan menyatakan bahwa "This limitation can undermine the reliability and validity of the assessment process" ketika merujuk pada variabilitas dalam peer assessment penulisan esai. Konsistensi antarpenilai menjadi tantangan utama dalam menjaga keadilan penilaian esai.
- Subjektivitas dalam Penilaian. Penilaian esai inherently bersifat subjektif karena melibatkan interpretasi konten, gaya penulisan, dan kualitas argumen. Tantangan ini memerlukan pendekatan sistematis untuk meminimalkan bias dan meningkatkan objektifitas.
- Solusi Teknis dan Metodologis Penelitian terkini mengidentifikasi beberapa solusi untuk mengatasi tantangan konsistensi:
    a. Penggunaan Rubrik dan Pelatihan Penilai Implementasi rubrik kriteria yang jelas dan pelatihan penilai secara berkelanjutan terbukti meningkatkan konsistensi penilaian esai (Rao & Banerjee, 2023).

b. Multiple Raters dan Model Hierarkis (Avcu, 2025) menunjukkan bahwa "models using HRM approach improves the reliability and validity of AES" (Automated Essay Scoring). Pendekatan Hierarchical Rater Models dapat meningkatkan reliabilitas baik dalam penilaian manual maupun otomatis.

2.3.2 Tes Esai Dan Hasil Pembelajaran

Tes esai adalah bentuk penilaian yang digunakan untuk mengevaluasi hasil belajar siswa, terutama keterampilan berpikir tingkat tinggi (HOTS) mereka. Tes ini mengharuskan siswa untuk membangun jawaban mereka sendiri daripada memilih dari serangkaian opsi, memungkinkan penilaian yang lebih dalam tentang kemampuan kognitif mereka, termasuk pemikiran kritis, analisis, dan argumentasi. Efektivitas dan karakteristik tes esai sering dipahami dalam kaitannya dengan proses kognitif yang ingin mereka ukur, seperti yang didefinisikan oleh taksonomi Bloom yang direvisi (Dabbagh & Safaei, 2019).

a. Definisi dan karakteristik

Perbedaan dari Tes Objektif: Tidak seperti tes objektif, yang memerlukan pemilihan jawaban yang benar, tes esai menuntut siswa menghasilkan dan mengatur tanggapan mereka sendiri. Format ini memungkinkan penilaian keterampilan kognitif kompleks di luar ingatan sederhana (Dabbagh & Safaei, 2019).

Mengukur Keterampilan Berpikir Tingkat Tinggi (HOTS) Pertanyaan esai sangat efektif dalam mengukur HOTS, yang meliputi menganalisis, mengevaluasi, dan menciptakan (Dabbagh & Safaei, 2019). Keterampilan ini melibatkan pemecahan materi menjadi bagian-bagian konstituen, membuat penilaian berdasarkan kriteria, dan menghasilkan ide atau produk baru. Sebuah studi pada buku pelajaran ELT menemukan bahwa buku pelajaran sekolah umum Iran, Prospect dan Visi, sangat menyukai Keterampilan Berpikir Tingkat Bawah (LOTS) seperti mengingat, memahami, dan menerapkan, dengan sedikit fokus pada HOTS (Dabbagh & Safaei, 2019). Sebaliknya, seri Four Corners yang diterbitkan secara internasional menunjukkan representasi yang lebih seimbang, meskipun masih dominan, dari kedua jenis keterampilan tersebut. Ini menunjukkan bahwa pertanyaan esai dirancang untuk menghasilkan HOTS. kurang umum dalam beberapa materi pendidikan.

b. Keuntungan dan kerugian

Kemampuan untuk Mengukur Keterampilan Kompleks: Tes esai unggul dalam mengukur kemampuan siswa untuk menganalisis, mensintesis, mengevaluasi, dan berpikir kritis. Mereka dapat mengungkapkan kapasitas siswa untuk mengatur pemikiran dan menyajikan argumen yang koheren.

Tantangan dalam Penilaian: Kerugian yang signifikan dari tes esai adalah subjektivitas yang terlibat dalam evaluasi mereka. Penilaian dapat dipengaruhi oleh bias evaluator, dan prosesnya seringkali memakan waktu dibandingkan dengan penilaian cepat dan objektif dari tes pilihan ganda. Keandalan penilaian adalah tantangan yang diketahui (Dabbagh & Safaei, 2019).

c. Tujuan dalam menilai hasil pembahasan

- Menilai Kemampuan Kritis dan Analitis: Tujuan utama menggunakan tes esai adalah untuk melampaui hafalan dan menilai kemampuan siswa untuk berpikir kritis, menganalisis informasi, dan membangun argumen yang beralasan. Mereka dirancang untuk mengevaluasi bagaimana siswa menerapkan pengetahuan dalam situasi baru atau asing (Dabbagh & Safaei, 2019).
- Relevansi dengan Kurikulum dan Prestasi: Tes esai harus selaras dengan tujuan kurikulum yang menekankan HOTS. Sebuah studi yang menganalisis buku teks ELT Iran mengungkapkan perbedaan yang signifikan antara tingkat kognitif yang dipromosikan dalam materi lokal versus internasional. Seri Prospect dan Vision lokal berfokus hampir secara eksklusif pada LOTS (92,94% tugas), mengabaikan keterampilan seperti menganalisis, mengevaluasi, dan menciptakan. SeriFour Corners, sementara juga berfokus pada LOTS (59,22%),

mendedikasikan porsi yang jauh lebih besar untuk HOTS (40,78%). (Dabbagh & Safaei, 2019)Ini menyoroti bagaimana desain alat penilaian, termasuk pertanyaan esai dalam buku teks, mencerminkan pencapaian pembelajaran yang dimaksudkan.

2.3.3 Teknik Dan Prosedur Penilaian Jawaban Esai

Penilaian esai manual mengikuti langkah: merancang rubrik, kalibrasi/percobaan, memberi skor, dan umpan balik formatif. Rubrik (holistik/analitik) meningkatkan transparansi tetapi reliabilitas memerlukan pelatihan penilai, penilaian ganda, moderasi, dan pemodelan perbedaan penilai.

a. Proses penilaian manual oleh guru

Penilaian manual biasanya dimulai dari perancangan instrumen (rubrik) yang menjabarkan kriteria dan tingkatan kualitas, lalu dilanjutkan dengan kalibrasi penilai, scoring, dan pemberian umpan balik kepada peserta didik. Literatur lapangan menekankan bahwa rubrik perlu divalidasi dan dicoba sebelum pemakaian untuk memastikan deskriptor jelas dan terukur (Viñas, 2022).

- Langkah membuat rubrik. Buat definisi konstruk, pilih kriteria esensial, rancang tingkat performa dengan deskripsi operasional, dan lakukan validasi ahli serta uji coba pada tulisan nyata untuk menilai kejelasan dan reliabilitas instrumen (Viñas, 2022).

- Langkah memberi skor. Lakukan sesi kalibrasi/coding antar-penilai menggunakan exemplar, skor setiap kriteria menurut deskriptor rubrik, dan catat skor serta komentar diagnostik untuk setiap siswa (Viñas, 2022).
- Langkah memberi umpan balik. Kombinasikan skor kriterial dengan komentar spesifik (kekuatan, area perbaikan, contoh perbaikan) sehingga umpan balik bersifat formatif dan dapat memandu revisi karya siswa (Arı, 2021).
- Faktor pengalaman penilai. Perbedaan pengalaman dan keahlian penilai memengaruhi keketatan/kelonggaran penilaian; studi empiris menemukan variasi signifikan pada dimensi lenient-severity antara penilai sehingga perlu dimodelkan atau dikalibrasi (Kernagaran & Abdullah, 2025).
- Bias dan kondisi penilai. Bias (misal preferensi gaya, implikasi budaya) dan kondisi penilai (kelelahan atau kebosanan akibat penandaan berulang) dapat menurunkan skor seiring waktu; penelitian menemukan peningkatan kebosanan berasosiasi dengan penurunan nilai meskipun rubrik diberikan (Viñas, 2022).
- Kriteria tidak konsisten. Beberapa deskriptor rubrik dapat bersifat samar sehingga menurunkan konsistensi antar-penilai; kajian lapangan melaporkan bahwa deskriptor yang ambigu

perlu direvisi untuk mengurangi perbedaan interpretasi (Uludag & McDonough, 2022).

b. Penggunaan rubrik penilaian esai

Rubrik hadir dalam bentuk utama holistik dan analitik, dan tiap tipe memiliki keunggulan serta keterbatasan yang berbeda dalam praktik kelas. Ringkasan perbandingan jenis rubrik berikut merangkum perbedaan fungsi, tujuan, dan implikasi reliabilitas berdasarkan kajian terkini (Viñas, 2022)(Asli et al., 2024)

- Manfaat rubrik untuk objektivitas dan transparansi. Rubrik mengurangi ketidakpastian tentang apa yang dinilai dan menyediakan deskripsi performa yang membuat penilaian lebih transparan bagi siswa dan penilai, sehingga meningkatkan keadilan dan pemahaman pembelajaran (Asli et al., 2024)
- Dampak pada pembelajaran. Rubrik analitik yang dibagikan kepada siswa mendukung evaluasi diri dan peer-assessment serta memfasilitasi perbaikan bertahap, meskipun penggunaan formatif membutuhkan desain deskriptor yang mendukung regulasi diri (Viñas, 2022).
- Contoh format rubrik dan kriteria. Instrumen yang divalidasi untuk esai argumentatif umumnya mencakup kriteria seperti organisasi/thesis, pengembangan/dukungan bukti, koherensi/logika argumen , pilihan leksikal dan tata bahasa , serta mekanik/presentasi, studi pengembangan rubrik

menampilkan format analitik 3–5 tingkat performa dengan deskriptor operasional untuk tiap kriteria (Baharudin et al., 2022) (Kloss & Burdiles, 2024).

- Panduan penulisan deskriptor. Tuliskan indikator yang dapat diamati/diukur, hindari bahasa normatif yang kabur, dan uji contoh esai exemplar untuk memastikan deskriptor dapat ditafsirkan konsisten oleh penilai (Asli et al., 2024).

c. Upaya meningkatkan reliabilitas penilaian esai

Meningkatkan reliabilitas inter-penilai memerlukan pendekatan prosedural (pelatihan, double scoring, moderasi) dan analitis/statistik (pemodelan perbedaan penilai). Bukti menunjukkan bahwa rubrik saja tidak cukup tanpa tindakan pendukung (Kernagaran & Abdullah, 2025).

- Pelatihan penilai. Sesi pelatihan yang mencakup kalibrasi pada exemplar, diskusi perbedaan skor, dan panduan pengguna rubrik meningkatkan literasi penilaian guru dan kualitas skala yang dikembangkan menurut studi pelatihan penilai di konteks universitas; penelitian melaporkan dampak positif pada kemampuan guru mengembangkan dan menggunakan skala analitik setelah pelatihan (Kvasova et al., n.d.).
- Penilaian ganda (double scoring). Menggunakan dua penilai independen per esai dan menggabungkan skor (mis. rata-rata atau aturan rekonsiliasi) meningkatkan kepercayaan skor;

validasi rubric sering melibatkan beberapa penilai sehingga koefisien reliabilitas internal atau ICC dapat dihitung untuk menilai konsistensi antar-penilai (Uludag & McDonough, 2022).

- Moderasi hasil penilaian. Moderasi (reconciliation meetings) untuk esai-yang-skor-menyimpang mengurangi perbedaan interpretasi rubrik dan menciptakan kesepakatan praktik penilaian yang lebih seragam (Viñas, 2022).
- Pendekatan statistik dan pemodelan. Model Many-Facet Rasch dan model IRT multidimensi dapat memodelkan efek penilai (lenient-severity) dan kualitas butir rubrik sehingga memungkinkan penyesuaian skor untuk bias penilai dan meningkatkan validitas skor akhir (Kernagaran & Abdullah, 2025).
- Langkah praktis operasional. Batasi jumlah esai yang dinilai berturut-turut untuk tiap penilai dan lakukan jeda untuk mengurangi efek kebosanan, penelitian eksperimen menunjukkan kebosanan meningkat seiring jumlah markah dan berkorelasi dengan penurunan skor, sehingga pengaturan beban penilaian adalah langkah mitigasi penting (Viñas, 2022).

2.4 Algoritma FastText

FastText adalah salah satu algoritma yang digunakan untuk menghasilkan representasi vektor kata dalam pemrosesan bahasa alami (NLP). Algoritma ini dikembangkan oleh Facebook AI Research pada tahun 2016 dan digunakan untuk

berbagai tugas NLP seperti klasifikasi teks, tagging entitas, dan analisis sentiment (Ainul Yaqin & Abdul Charis Fauzan, 2025). Selain itu fasttext merupakan pustaka open source yang dirancang untuk mempelajari representasi kata (word/sentence vectors) atau sebuah subword atau caracter dimana setiap kata direpresentasikan sebagai kumpulan potongan karakter (Hartoko & Leong, 2023) sehingga fasttext lebih tangguh menghadapi kata baru (OOV), variasi morfologi, dan salah eja serta pelatihan dan inferensinya sangat cepat di CPU biasa dengan akurasi yang kompetitif sebagai baseline klasifikasi teks. Fasstext bekerja dengan cara merepresentasikan kata sebagai jumlah vektor n-gram karakternya misalnya kata playing terdiri dari potongan seperti pla, play, layi sehingga mengurangi OOV dan membantu bahasa dengan morfologi yang kaya (Sutriawan et al., 2025). Fasttext mengadaptasi cara CBOW/skip-gram dengan tambahan *negative sampling* atau *hierarchical softmax* untuk efisiensi pelatihan. Selain itu dalam kasus klasifikasi fasttext menggunakan model linear dengan fitur n-gram kata, tricks sederhana dan aproksimasi softmax yang membuatnya sangat cepat (Hartoko & Leong, 2023).

Fasttext banyak digunakan sebagian peneliti karena memiliki beberapa kelebihan antara lain adalah kecepatan pelatihan pada skala besar menggunakan CPU, kemampuan mengatasi kata baru melalui representasi subword, serta ketersediaan vektor pralatih (*pre-trained word vectors*) untuk lebih dari 150 bahasa, termasuk Bahasa Indonesia. Model pralatih tersebut dibangun menggunakan korpus Common Crawl dan Wikipedia, sehingga memudahkan peneliti untuk langsung menggunakannya sebagai fitur dalam berbagai tugas NLP.

Disisi lain, FastText juga bersifat *open source* dan dapat diintegrasikan dengan mudah melalui antarmuka baris perintah (CLI) maupun pustaka Python/C++. Selain itu fasttext juga mampu mengatasi masalah kata yang jarang ditemukan (out of vocabulary words) karena memperhitungkan subword yang merepresentasikan vektor kata yang berkualitas meskipun dalam dataset yang besar (Ainul Yaqin & Abdul Charis Fauzan, 2025)

Tetapi FastText memiliki beberapa keterbatasan yaitu model ini hanya mempertimbangkan konteks lokal dalam bentuk *bag-of-n-grams*, sehingga tidak mampu menangkap makna kontekstual yang kompleks sebagaimana model modern seperti BERT atau ELMo. Selain itu, FastText tidak secara eksplisit memodelkan struktur sintaksis atau hubungan antar kalimat, sehingga performanya sangat bergantung pada pemilihan *n-gram* dan kualitas data yang digunakan. Meskipun demikian, karena kecepatan dan efisiensinya, FastText masih banyak digunakan untuk aplikasi NLP di lingkungan dengan sumber daya terbatas. Berikut pseudocode sederhana pada algoritma fasttext :

```
function trainFastText(dataset: List[str], num_epochs: int,
learning_rate: float) -> Model:
    model = initializeModel() // inisialisasi model
    for epoch in range(num_epochs):
        for text in dataset:
            subwords = generateSubwords(text) // menghasilkan
subword dari teks
            for subword in subwords:

                loss = computeLoss(model, subword) // menghitung
nilai loss
                updateWeights(model, subword, loss, learning_rate)
// update bobot model
    return model
```

Gambar 2. 2 Pseudocode Algoritma FastText

2.5 Algoritma LSTM

Long Short-Term Memory (LSTM) adalah arsitektur jaringan saraf berulang (RNN) yang diperkenalkan oleh Hochreiter & Schmidhuber untuk mengatasi masalah *vanishing/exploding gradients*, dengan memperkenalkan sel memori dan tiga gerbang (input, output, dan forget) yang mengatur aliran informasi sehingga model mampu mempelajari ketergantungan jangka panjang pada data berurutan (mis. teks, suara, deret waktu) (Malashin et al., 2024; Waqas & Humphries, 2024). Berikut dapat dilihat pseudocode dari algoritma LSTM :

```
procedure LSTM(input_sequence, initial_hidden_state,
initial_cell_state, W, U, b, V, c, output_sequence)
    // input_sequence: sequence of input vectors
    // initial_hidden_state: initial hidden state vector
    // initial_cell_state: initial cell state vector
    // W, U, b, V, c: learnable parameters
    // output_sequence: empty sequence to store output vectors

    // initialize hidden state and cell state
    h = initial_hidden_state
    c = initial_cell_state

    // iterate through input sequence
    for i = 1 to length(input_sequence) do
        // calculate gates
        forget_gate = sigmoid(W_f * input_sequence[i] + U_f * h +
b_f)
        input_gate = sigmoid(W_i * input_sequence[i] + U_i * h +
b_i)
        output_gate = sigmoid(W_o * input_sequence[i] + U_o * h +
b_o)
        candidate_cell_state = tanh(W_c * input_sequence[i] + U_c *
h + b_c)

        // update cell state
        c = forget_gate * c + input_gate * candidate_cell_state

        // update hidden state
        h = output_gate * tanh(c)

        // calculate output vector
        output_sequence[i] = softmax(V * h + c)
    end for
end procedure
```

Gambar 2. 3 Pseudocode Algoritma LSTM

Dari Gambar 3.2 dapat dilihat bahwa algoritma LSTM menerima urutan vektor masukan (input_sequence) bersama dengan vektor keadaan awal yang terdiri dari initial_hidden_state dan initial_cell_state (Ainul Yaqin & Abdul Charis Fauzan, 2025). Selain itu, algoritma ini juga menggunakan parameter yang dapat dipelajari, meliputi bobot (W, U), bias (b), serta matriks (V, c) dan hasil keluaran dari proses ini berupa urutan vektor keluaran (output_sequence).

Selama proses iterasi pada setiap vektor input, LSTM menghitung sejumlah gerbang yang meliputi forget gate, input gate dan output gate, serta kandidat keadaan sel (candidate cell state). Perhitungan ini dilakukan dengan menggunakan fungsi sigmoid untuk setiap gerbang dan fungsi tangen hiperbolik (tanh) untuk menentukan nilai kandidat keadaan sel. Kemudian, keadaan sel (cell state) diperbarui dengan mengombinasikan hasil dari forget gate dan input gate terhadap kandidat keadaan sel. Setelah itu, hidden state (h) diperbarui menggunakan output gate dan fungsi tanh dari keadaan sel yang telah diperbarui.

Untuk setiap langkah waktu, vektor keluaran dihasilkan dengan menerapkan fungsi softmax pada hasil perkalian antara matriks bobot keluaran (V) dan hidden state (h), lalu menambahkan bias (c). Hasil tersebut kemudian disimpan sebagai bagian dari urutan keluaran. Algoritma ini menunjukkan bagaimana LSTM mampu memproses urutan data sambil mempertahankan informasi jangka panjang melalui cell state, sehingga sangat efektif digunakan pada tugas-tugas berbasis urutan, seperti pemrosesan bahasa alami dan prediksi deret waktu.

Dalam praktiknya, LSTM dilatih dengan *backpropagation through time* dan hadir dalam berbagai varian seperti (dua arah), stacked LSTM ((bertumpuk),

peephole LSTM, serta pengayaan dengan mekanisme atensi. Kelebihan yang dimiliki LSTM adalah unggul untuk *sequence modeling* dengan ketergantungan menengah - panjang, bekerja baik pada data berukuran kecil - menengah dan skenario inferensi berbasis CPU dengan latensi ketat. Selain itu LSTM juga memiliki keterbatasan yaitu pelatihan yang kurang paralel (dibanding Transformer) sensitivitas hiperparameter, dan tantangan menangkap dependensi yang sangat panjang.

BAB III

METODOLOGI PENELITIAN

Pada bagian ini menjelaskan langkah-langkah sistematis dalam menyelesaikan permasalahan prediksi skor jawaban esai menggunakan Fasttext dan algoritma Long Short Term Memory (LSTM). Seluruh tahapan yang dilakukan pada penelitian ini dijabarkan dalam desain peneltian, kerangka model dan desain eksperimen.

3.1 Deskripsi Data

Data penelitian ini berupa korpus jawaban esai siswa pada mata pelajaran Bahasa Inggris tingkat Sekolah Menengah Pertama (SMP). Korpus disusun sebagai tabel terstruktur berisi 500 observasi yang berasal dari 100 siswa, masing-masing mengerjakan lima butir soal esai (100 × 5 = 500). Setiap observasi merepresentasikan satu pasangan unik “siswa dengan butir soal”, yakni satu respons esai yang ditulis seorang siswa untuk satu pertanyaan terbuka dalam Bahasa Inggris yang menuntut konstruksi kalimat atau paragraf. Dengan demikian, relasi datanya bersifat one-to-many: satu siswa menghasilkan lima respons.

Skema tabel memuat lima kolom utama: NISN sebagai pengenal unik siswa (numerik atau string terformat), nama lengkap siswa sebagaimana tercatat di administrasi sekolah (string), teks butir pertanyaan esai Bahasa Inggris, teks

jawaban esai yang dapat terdiri atas satu hingga beberapa paragraf, serta skor numerik hasil penilaian guru. Penamaan kolom mengikuti sumber asli antara lain. “Soal esai” dan “jawaban esai”. Kunci konseptual setiap baris adalah gabungan identitas siswa dan penanda butir (misalnya indeks soal), sehingga tiap baris dapat dipandang unik pada tingkat pasangan (siswa, soal).

Konten jawaban menunjukkan variasi panjang yang tinggi akibat sifat tugas yang bebas mulai dari satu kalimat hingga beberapa paragraf. Mengingat konteks SMP, karakteristik kebahasaan yang muncul meliputi kemungkinan salah eja, tata bahasa sederhana, dan sesekali penyisipan kata/ungkapan berbahasa Indonesia (code-switching). Seluruh respons berada dalam domain yang homogen karena kelima prompt diberikan seragam kepada semua siswa. Label penilaian berupa skor numerik yang diberikan oleh guru sesuai rubrik esai (misalnya aspek isi, organisasi, tata bahasa, kosakata, dan mekanika). Skala yang digunakan adalah skala tunggal yaitu 0–100. Dalam pemodelan, skor diperlakukan sebagai variabel target untuk pendekatan regresi (prediksi nilai kontinu).

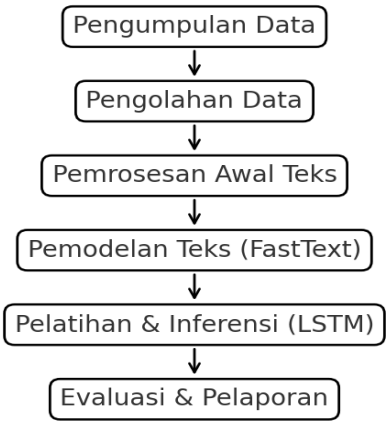
Aspek kualitas dan etika penanganan data diperhatikan secara khusus. Kolom NISN dan nama siswa termasuk data identitas pribadi sehingga untuk keperluan analisis dan publikasi akan dilakukan pseudonimisasi misalnya hashing NISN dan penggantian nama dengan kode. Pemeriksaan konsistensi dilakukan terhadap duplikasi pasangan (siswa, soal), validitas format NISN, dan kecocokan jumlah baris (harus 500). Nilai kosong pada kolom wajib (NISN, soal, jawaban, atau skor) akan diidentifikasi; baris yang tidak lengkap dapat dikeluarkan sesuai protokol pra-pemrosesan, sedangkan imputasi untuk skor tidak dilakukan.

Implikasi terhadap rancangan analisis mencakup strategi pembagian data dan pra-pemrosesan. Karena terdapat ketergantungan intra-siswa (lima respons per siswa), pemisahan train/validation/test menggunakan skema berbasis kelompok siswa guna mencegah kebocoran data lintas siswa.

Keterbatasan utama korpus ini adalah konteks tunggal hanya Bahasa Inggris di jenjang SMP sehingga generalisasi ke jenjang atau mata pelajaran lain memerlukan kehati-hatian. Jika penilaian dilakukan oleh satu penilai, potensi bias antar-penilai tidak dapat diestimasi. Selain itu, sifat "teks pelajar" awal (beginner learner language) membawa noise linguistik yang menantang namun relevan bagi skenario penilaian otomatis. Secara keseluruhan, dataset ini menyediakan 500 respons esai berlabel dengan struktur sederhana namun memadai sebagai dasar eksperimen pemodelan penilaian otomatis, sambil tetap menjaga kepatuhan etika dan validitas metodologis.

3.2 Desain Penelitian

Pada penelitian ini melakukan penelitian yang bersifat kuantitatif dimana tujuan yang dihasilkan berupa angka. Angka tersebut didapatkan dari hasil prediksi. Terdapat beberapa langkah atau tahapan yang dilakukan pada penelitian ini untuk menyelesaikan permasalahan prediksi skor jawaban esai menggunakan Fasttext dan Algoritma Long Short Term Memory (LSTM).



Gambar 3. 1 Desain Penelitian

Pada Gambar 3.1 ditunjukkan tahapan penelitian yang meliputi pengumpulan data, pengolahan data, pemrosesan awal data, pemodelan teks, algoritma LSTM, dan evaluasi model. Pengumpulan data merupakan tahap awal dalam penelitian ini untuk memastikan kesiapan data. Setelah data didapatkan, tahap selanjutnya adalah melakukan pengolahan data dengan tujuan untuk menentukan fitur mana yang akan digunakan. Selanjutnya, dilakukan pemrosesan awal teks. Untuk memastikan bahwa data dapat dijadikan sebagai input pada tahap algoritma LSTM, data teks tersebut harus diubah menjadi representasi angka dengan menggunakan fasttext yang dilakukan pada tahap pemodelan teks. Pada tahap algoritma LSTM merupakan tahapan untuk melatih model yang selanjutnya model tersebut akan dievaluasi kinerjanya pada tahap evaluasi model.

3.2.1 Pengumpulan Data.

Data yang digunakan dalam penelitian ini dikumpulkan dari SMP Negeri 6 Singosari untuk mata pelajaran Bahasa Inggris. Pemilihan sekolah ini didasarkan

pada pertimbangan aksesibilitas data, kualitas dokumentasi penilaian, serta representativitas siswa sebagai subjek penelitian. Pengumpulan data dilakukan melalui proses pengajuan izin penelitian kepada pihak sekolah, diikuti dengan koordinasi bersama guru Bahasa Inggris yang bersangkutan untuk memperoleh lembar jawaban siswa beserta rubrik penilaian yang digunakan.

Total data yang digunakan dalam penelitian ini berjumlah 500 data, yang terdiri dari jawaban 100 siswa untuk 5 butir soal esai. Setiap data merepresentasikan satu pasangan soal-jawaban dari seorang siswa, disertai dengan skor yang diberikan oleh guru sebagai ground truth. Kelima butir soal dirancang dengan tingkat kesulitan yang bervariasi dan mencakup tema-tema yang berbeda, antara lain hobi dan minat, deskripsi orang atau tempat, pengalaman pribadi, pendapat dan argumentasi, serta narasi cerita.

Skor yang diberikan guru berkisar dari 0 hingga 100 dan didasarkan pada rubrik penilaian esai yang mencakup aspek isi (content), tata bahasa (grammar), kosakata (vocabulary), koherensi (coherence), dan mekanik penulisan (mechanics). Penilaian dilakukan oleh dua orang guru secara independen, dan skor final merupakan rata-rata dari keduanya untuk memastikan reliabilitas dan objektivitas penilaian.

Pembagian dataset dilakukan dengan rasio 70:15:15 untuk data pelatihan, validasi, dan pengujian secara berturut-turut. Dengan demikian, 350 data digunakan untuk pelatihan model, 75 data untuk validasi selama proses pelatihan, dan 75 data untuk pengujian akhir performa model. Stratifikasi pembagian data mempertimbangkan distribusi skor agar ketiga subset memiliki representasi yang

seimbang dari seluruh rentang skor. Pada tabel 3.1 ditunjukkan contoh struktur dataset yang akan digunakan pada penelitian ini.

Tabel 3. 1 Dataset Jawaban Esai

NISN	Soal	Jawaban	Skor
1234567890	Describe your daily routine in simple present tense.	I wake up at 5 a.m., then I pray, take a shower, and go to school by bike.	85
1234567891	Explain the difference between 'since' and 'for'.	Since is used for a point in time; for is used for a duration of time.	90
1234567892	Write a short paragraph about your hobby.	I love playing badminton on weekends with my friends at the community hall.	80

3.2.2 Pengolahan Data

Pada tahap ini akan melakukan pengolahan data yang bertujuan untuk menghasilkan data yang siap untuk diproses pada tahap selanjutnya. Tahap pengolahann data dilakukan dengan beberapa Langkah yang meliputi penghapusan identitas data dalam hal ini adalah kolom NISN karena tidak memiliki pengaruh terhadap target. Langkah lainnya adalah melakukan penggabungan kolom soal dengan kolom jawaban untuk menjaga keterkaitan konteks, serta menabahkan penanda khusus “<startanswer>” dan “<endanswer>” agar model dapat membedakan antara teks soal dan teks jawaban. Contoh hasil proses pengolahan data ditunjukkan pada tabel 3.2.

Tabel 3. 2 Contoh Hasil Pengolahan Data

Teks_Gabungan	Skor
soal: describe your daily routine in simple present tense. <startanswer> i wake up at 5 a.m., then i pray, take a shower, and go to school by bike. <endanswer>	85
soal: explain the difference between 'since' and 'for'. <startanswer> since is used for a point in time; for is used for a duration of time. <endanswer>	90
soal: write a short paragraph about your hobby. <startanswer> i love playing badminton on weekends with my friends at the community hall. <endanswer>	80

3.2.3 Pemrosesan Awal Teks

Pada tahap pemrosesan awal teks digunakan untuk melakukan pembersihan dan menormalkan data. Tahap pemrosesan awal teks yang terdapat pada penelitian ini meliputi proses case folding dan tokenizer. Proses case folding merupakan proses yang digunakan untuk mengubah seluruh teks menjadi huruf kecil. Proses ini dilakukan karena kasus yang terdapat pada penelitian ini,penulisan huruf kapital tidak memiliki pengaruh terhadap target. Proses tokenizer merupakan proses yang digunakan untuk memecah data teks menjadi Kumpulan token. Proses ini dilakukan karena pada tahap pemodelan teks membutuhkan input yang berupa token. Pada tabel 3.3 ditunjukkan contoh hasil dari tahap pemrosesan awal teks.

Tabel 3. 3 Contoh Hasil dari Pemrosesan Awal Teks

Kalimat_Asli	Lowercase	Token
I wake up at 5 a.m.	i wake up at 5 a.m.	[“i”,”wake”,”up”,”at”,”5”,”a.m”]
Since is used for a point in time	since is used for a point in time	[“since”,’is”,”used”,”for”,”a”,”point”, “in”,”time”]

3.2.4 Pemodelan Teks

Pemodelan teks adalah tahap penting dalam Natural Language Processing (NLP) yang bertujuan mengubah kata, kalimat, atau dokumen menjadi vektor angka yang mewakili makna dari teks tersebut. Komputer tidak bisa memahami kata secara langsung, sehingga teks perlu direpresentasikan dalam bentuk angka agar model dapat mengenali pola, kemiripan makna, dan hubungan antar kata. Dalam penelitian ini, metode pemodelan teks yang digunakan adalah FastText , karena memiliki keunggulan dalam memahami struktur kata melalui pendekatan subword modeling , yaitu memecah kata menjadi potongan-potongan kecil (subword atau n-gram huruf). Misalnya, kata *"playing"* akan dipecah menjadi potongan seperti "pla", "lay", dan "ing". Setiap potongan kecil ini memiliki vektor tersendiri yang dilatih selama proses pembelajaran. Dengan cara ini, FastText mampu mengenali makna kata berdasarkan bagian-bagian pembentuknya, sehingga model tetap dapat menghasilkan vektor yang bermakna untuk kata-kata yang belum pernah muncul sebelumnya (out-of-vocabulary), salah ejaan, atau bentuk turunan dari kata dasar seperti "play", "plays", dan "played" yang semuanya memiliki kemiripan struktur.

Selain itu FastText juga efisien dan mudah digunakan dibandingkan model lain yang lebih kompleks. Model ini mampu menghasilkan representasi kata yang stabil dan akurat meskipun data latihnya tidak terlalu besar. Proses pemodelan tidak hanya menggunakan model bawaan FastText, tetapi juga dilakukan pelatihan ulang (retraining) menggunakan korpus gabungan yang berisi teks dari soal dan jawaban. Tujuannya agar model FastText memahami konteks dan kosa

kata yang khas dari domain data penelitian, bukan hanya pengetahuan umum dari model pra-latih yang biasanya dilatih pada korpus besar seperti Wikipedia. Dengan pelatihan ulang ini, vektor kata yang dihasilkan menjadi lebih relevan terhadap bahasa dan istilah yang muncul pada data soal-jawaban.

Setiap kata dalam model FastText direpresentasikan dalam bentuk vektor berdimensi 300 , artinya satu kata diterjemahkan menjadi deretan 300 angka yang masing-masing mewakili aspek makna tertentu dari kata tersebut. Ukuran 300 dipilih karena seimbang antara ketepatan representasi dan efisiensi komputasi tidak terlalu kecil sehingga kehilangan makna, dan tidak terlalu besar sehingga membebani sistem. Setelah semua kata diubah menjadi vektor, kalimat atau dokumen bisa direpresentasikan sebagai rata-rata dari seluruh vektor kata di dalamnya. Representasi inilah yang kemudian digunakan sebagai masukan (input) untuk algoritma pembelajaran mesin seperti klasifikasi teks dalam penelitian ini adalah algoritma LSTM.

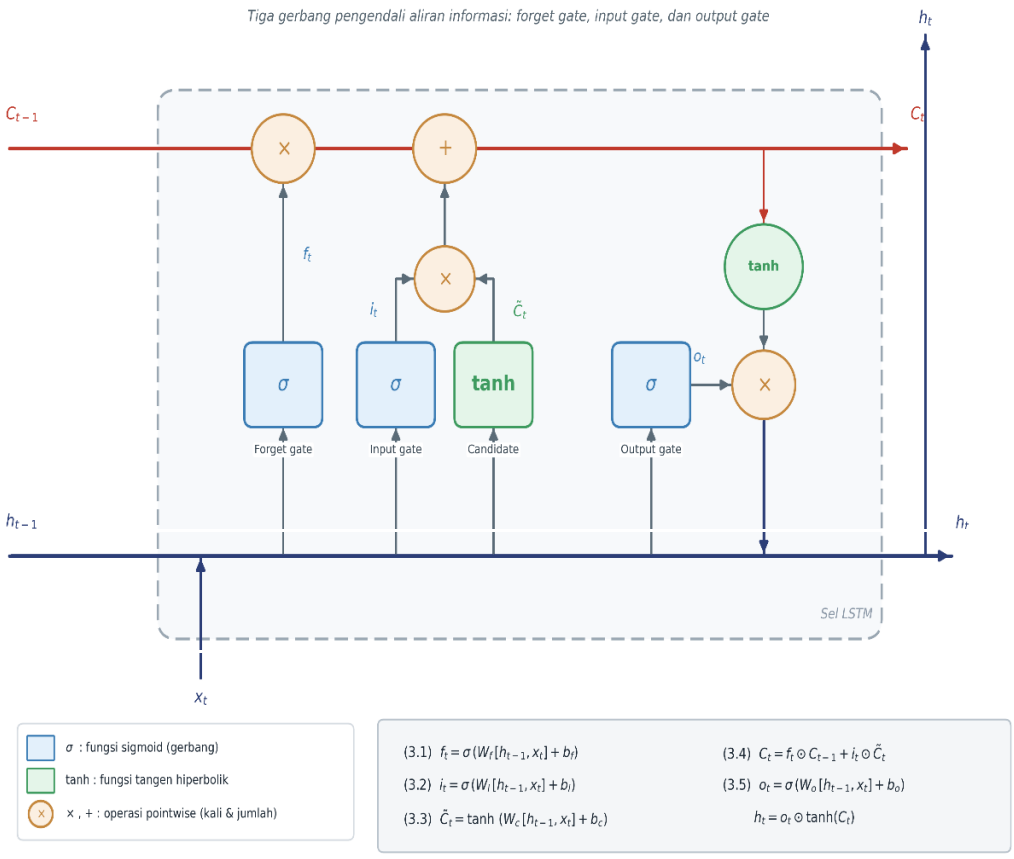
Proses pemodelan teks berbasis FastText membantu sistem memahami hubungan semantik antar kata dengan cara yang lebih fleksibel dan tahan terhadap variasi penulisan. Keunggulan lain dari pendekatan ini adalah kemampuannya menangani kata baru atau langka tanpa kehilangan makna karena pembelajaran berbasis potongan huruf. Pada Tabel 3.4 ditunjukkan contoh vektor angka dari hasil pemodelan teks menggunakan FastText.

Tabel 3. 4 Contoh Hasil Pemodelan Teks Menggunakan FastText

Kalimat asli	Token	0	1	2	3	4	5	6	7	8	9	10	...	299
I wake up at 5 a.m	i	-0.3548609	0.2955139	0.1963196	0.7158919	-0.31943408	-0.4364896	-0.11823853	-0.83765036	0.74644786	-0.4483518	0.259177		-16.986.039
	wake	-0.0075262915	-0.47992176	0.49592185	0.07883707	-0.17747076	0.16802631	0.5330683	0.07896935	0.029734433	0.430934	0.3192784		-0.4507001
	up	0.07651632	-10.560.133	-0.24249624	0.80333847	-0.19644931	-0.8855531	-0.08250305	0.66203797	0.08986515	-0.5293385	-0.49841112		0.16357131
	at	-0.73948175	0.17347741	0.9074367	0.20751026	-10.651.312	0.16806357	-0.38886407	-0.12840751	-0.7238136	-0.6618754	-0.14282535		-0.2703793
	5	0.18550825	0.36208513	0.86651725	-0.4172758	-10.430.965	-0.08178815	-0.9077975	0.6883542	-0.9243837	-0.25788945	0.3492676		-0.43930697
	a.m	-0.019528585	-0.8160104	0.1655495	0.6346028	0.0127844745	0.026755307	-0.59945816	-0.5700798	-0.077724725	-0.31121922	-0.08731777		0.35331726

3.2.5 Arsitektur Model LSTM

Long Short-Term Memory (LSTM) merupakan varian dari Recurrent Neural Network (RNN) yang dirancang khusus untuk mengatasi masalah vanishing gradient yang umum terjadi pada RNN standar. LSTM diperkenalkan oleh Hochreiter dan Schmidhuber pada tahun 1997 dan telah menjadi salah satu arsitektur paling berpengaruh dalam pemrosesan data sekuensial, termasuk teks dan rangkaian waktu (time series). Pada gambar 3.2 ditunjukkan visualisasi model LSTM (Hochreiter & Schmidhuber, 1997).



Gambar 3. 2 Arsitektur Model LSTM

Komponen utama sebuah sel LSTM adalah sistem tiga gerbang (gates) yang mengontrol aliran informasi: forget gate, input gate, dan output gate. Forget gate menentukan informasi mana dari cell state sebelumnya yang akan dipertahankan atau dilupakan. Input gate menentukan informasi baru mana yang akan ditambahkan ke cell state. Output gate menentukan bagian mana dari cell state yang akan menjadi output (hidden state) dari sel LSTM tersebut. Cell state bertindak sebagai memori jangka panjang yang mengalir sepanjang urutan dengan modifikasi yang relatif kecil, sehingga memungkinkan informasi dari langkah-langkah yang jauh tetap dapat diakses.

Secara matematis, mekanisme gerbang LSTM dapat diformulasikan sebagai berikut. Forget gate dihitung dengan menggunakan persamaan 3.1

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \quad (3.1)$$

di mana σ adalah fungsi sigmoid, h_{t-1} adalah hidden state sebelumnya, x_t adalah input pada waktu t, W_f adalah matriks bobot, dan b_f adalah bias.

Sedangkan input gate dihitung dengan persamaan 3.2

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (3.2)$$

dan kandidat cell state baru dihitung dengan persamaan 3.3

$$\tilde{c}_t = tanh(W_c[h_{t-1}, x_t] + b_c) \quad (3.3)$$

Kemudian Cell state diperbarui dengan persamaan 3.4

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (3.4)$$

dan output gate serta hidden state akhir dihitung dengan menggunakan persamaan

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (3.5\ )$$

Model LSTM yang digunakan dalam penelitian ini menggunakan arsitektur multi-layer yang terdiri dari beberapa layer berurutan. Layer pertama adalah Embedding Layer yang menerima representasi FastText dengan dimensi 300. Layer ini diatur sebagai non-trainable (weights tidak diperbarui selama pelatihan) untuk mempertahankan representasi semantik yang telah dipelajari dari model FastText pre-trained. Layer kedua adalah LSTM Layer dengan 128 unit (sesuai hasil hyperparameter tuning). Setelah LSTM Layer, diterapkan Dropout Layer dengan rate 0.3 untuk mencegah overfitting. Layer berikutnya adalah Dense Layer dengan 64 neuron dan fungsi aktivasi ReLU yang berfungsi sebagai lapisan klasifikasi non-linear. Output layer adalah Dense Layer dengan satu neuron dan fungsi aktivasi linear untuk menghasilkan prediksi skor dalam rentang kontinu.

Tabel 3. 5 Spesifikasi Arsitektur Model LSTM

Layer	Tipe	Konfigurasi	Output Shape	Parameter
1	Embedding (FastText)	300 dim, non-trainable, seq_len=150	(None, 150, 300)	0 (frozen)
2	LSTM	128 units, return_sequences=False	(None, 128)	219,648
3	Dropout	rate=0.3	(None, 128)	0
4	Dense	64 units, aktivasi ReLU	(None, 64)	8,256
5	Dropout	rate=0.2	(None, 64)	0
6	Dense (Output)	1 unit, aktivasi Linear	(None, 1)	65
Total	—	—	—	228,969 parameter

3.2.6 Arsitektur Model Bi-LSTM

Bidirectional LSTM (Bi-LSTM) merupakan pengembangan dari arsitektur LSTM standar yang memproses urutan input dari dua arah secara bersamaan: arah maju (forward) dan arah mundur (backward). Konsep ini pertama kali diperkenalkan oleh Schuster dan Paliwal pada tahun 1997 untuk jaringan saraf berulang, dan kemudian diadaptasi untuk arsitektur LSTM.

Dalam LSTM unidireksional, representasi suatu kata pada posisi t hanya dipengaruhi oleh kata-kata yang muncul sebelumnya (posisi 1 hingga t-1). Hal ini merupakan keterbatasan dalam pemrosesan teks karena makna sebuah kata sering kali dipengaruhi oleh konteks yang muncul setelahnya. Bi-LSTM mengatasi keterbatasan ini dengan menjalankan dua proses LSTM secara paralel: satu LSTM memproses urutan dari kiri ke kanan (maju), sementara LSTM lainnya memproses dari kanan ke kiri (mundur). Hidden state akhir dari kedua arah kemudian digabungkan (biasanya melalui concatenation) untuk menghasilkan representasi final yang mencakup konteks dari kedua arah.

Dalam konteks penilaian esai, kemampuan Bi-LSTM untuk menangkap konteks dari kedua arah sangat berguna. Kata-kata di bagian akhir jawaban dapat mempengaruhi interpretasi kata-kata di awal, dan sebaliknya. Misalnya, dalam kalimat "I enjoy reading, especially academic books about science", kata "science" di bagian akhir memberikan konteks yang memperjelas jenis buku yang dimaksud pada awal kalimat. LSTM unidireksional tidak dapat memanfaatkan informasi ini ketika memproses kata "reading".

Karena terdiri dari dua LSTM yang berjalan paralel, jumlah parameter dalam model Bi-LSTM kira-kira dua kali lipat dari model LSTM dengan jumlah unit yang sama. Dalam penelitian ini, Bi-LSTM menggunakan 128 unit per arah, sehingga total 256 unit output dari layer Bi-LSTM sebelum penggabungan. Model Bi-LSTM menggunakan arsitektur dan hyperparameter yang sama dengan model LSTM untuk mendukung kondisi perbandingan yang adil dan setara.

Tabel 3. 6 Spesifikasi Arsitektur Model Bi-LSTM

Layer	Tipe	Konfigurasi	Output Shape	Parameter
1	Embedding (FastText)	300 dim, non-trainable, seq_len=150	(None, 150, 300)	0 (frozen)
2	Bi-LSTM	128 units × 2 arah, concat	(None, 256)	439,296
3	Dropout	rate=0.3	(None, 256)	0
4	Dense	64 units, aktivasi ReLU	(None, 64)	16,448
5	Dropout	rate=0.2	(None, 64)	0
6	Dense (Output)	1 unit, aktivasi Linear	(None, 1)	65
Total	—	—	—	455,809 parameter

3.2.7 Evaluasi Model

Evaluasi performa model dilakukan menggunakan metrik Mean Absolute Percentage Error (MAPE). Pemilihan MAPE sebagai metrik evaluasi didasarkan pada kemudahan interpretasinya dalam konteks penilaian esai, karena MAPE mengukur rata-rata persentase selisih antara skor prediksi dan skor aktual relatif terhadap nilai skor aktual. MAPE dihitung menggunakan persamaan 3.6 .

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_{aktual} - y_{prediksi}|}{y_{aktual}} \times 100\% \quad (3.6)$$

di mana n adalah jumlah data uji, y_aktual adalah skor aktual yang diberikan guru, dan y_prediksi adalah skor hasil prediksi model. Nilai MAPE yang lebih rendah menunjukkan performa model yang lebih baik.

Tabel 3. 7 Interpretasi Nilai MAPE

Rentang MAPE	Kategori Performa	Deskripsi
< 10%	Sangat Baik	Prediksi sangat akurat, cocok untuk implementasi produksi
10% – 20%	Baik	Prediksi cukup akurat untuk sebagian besar aplikasi praktis

Rentang MAPE	Kategori Performa	Deskripsi
20% – 50%	Dapat Diterima	Prediksi memiliki akurasi yang cukup dengan penyimpangan yang masih dapat ditoleransi
> 50%	Buruk	Prediksi tidak akurat, model memerlukan perbaikan signifikan

3.3 Desain Eksperimen.

Pada tahap ini dilakukan perancangan eksperimen. Eksperimen digunakan untuk merancang beberapa strategi penyelesaian pada permasalahan prediksi skor jawaban esai menggunakan fasttext dan algoritma LSTM. Adapun strategi yang dilakukan pada eksperimen adalah sebagai berikut :

1. Pengujian *hyperparameter tuning*

   Pengujian hyperparameter dilakukan secara bertahap (sequential), yaitu setiap hyperparameter dioptimalkan satu per satu dengan mempertahankan hyperparameter lainnya pada nilai default, kemudian nilai terbaik yang diperoleh digunakan pada tahap pengujian berikutnya. Terdapat empat hyperparameter yang diuji, yaitu jumlah unit hidden layer, learning rate, batch size, dan dropout rate. Setiap konfigurasi dilatih menggunakan optimizer Adam dengan fungsi loss Mean Squared Error (MSE), maksimum 100 epoch, dan mekanisme early stopping dengan patience 10 epoch. Setiap konfigurasi diuji sebanyak lima kali dengan inisialisasi bobot acak yang berbeda, dan nilai Mean Absolute Percentage Error (MAPE) rata-rata digunakan sebagai metrik pemilihan konfigurasi terbaik. Tabel 3.8 menunjukkan nilai-nilai yang diuji untuk setiap hyperparameter.

Tabel 3. 8 Nilai Hyperparameter yang Diuji

No.	Hyperparameter	Nilai yang Diuji
1.	Jumlah Unit Hidden Layer	32, 64, 128, 256, 512
2.	Learning Rate	0.0001, 0.0005, 0.001, 0.005, 0.01
3.	Batch Size	16, 32, 64, 128
4.	Dropout Rate	0.1, 0.2, 0.3, 0.4, 0.5

2. Pengujian berdasarkan karakteristik data

Pengujian ini dilakukan dengan mengelompokkan berdasarkan karakteristik data. Terdapat beberapa pengujian yang akan dilakukan antara lain :

a. Pengujian berdasarkan kelompok panjang kata.

   Pengujian ini dilakukan dengan membagi data menjadi 2 kelompok yaitu kelompok yang memiliki kalimat pendek dan kalimat yang panjang. Tujuan dari penelitian ini digunakan untuk mengetahui keandalan dari model untuk memodelkan data yang memiliki kalimat pendek dan kalimat panjang. Hal ini dilakukan karena jawaban esai yang diberikan siswa memiliki variasi panjang kalimat yang berbeda. Hal tersebut juga menunjukkan tingkat kepahaman siswa.

b. Pengujian berdasarkan pengelompokkan skor.

   Pengujian ini dilakukan data menjadi 3 kelompok yakni kelompok skor diatas KKM, skor KKM, dan skor di bawah KKM. Tujuan dari pengujian ini dilakukan untuk mendeteksi seberapa handal model dalam mempelajari data dengan beberapa kelompok tersebut.

c. Pengujian dengan pembagian data berdasarkan siswa vs soal.

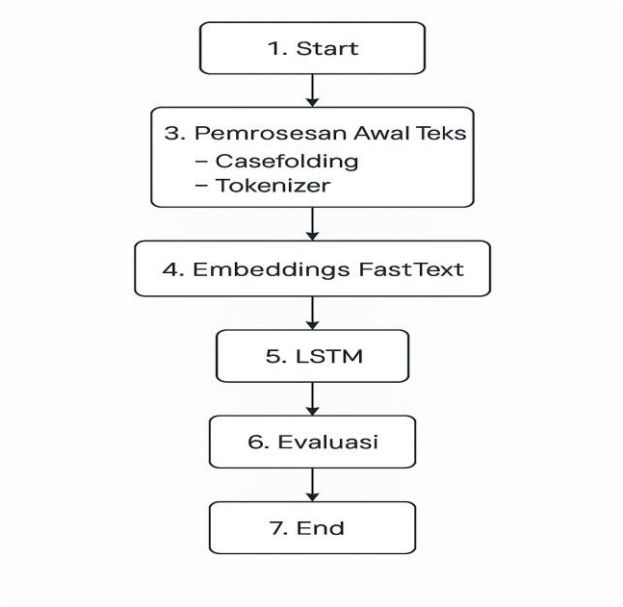
   Data yang digunakan pada penelitian ini sejumlah 500 yang terdiri dari 100 siswa x 5 butir soal. Setiap siswa memiliki pola yang berbeda dalam

menjawab baik soal baik secara penyusunan kalimat dan lainnya. Sehingga dalam pengujian ini diterapkan pada proses pembagian data dimana pembagian data berdasarkan siswa yaitu 100 siswa.

Untuk menjalankan eksperimen dalam penelitian ini mengunakan IDE google colab, bahasa pemrograman python dan library tensorflow dan keras untuk algoritma LSTM, gensim untuk fasttext, scikit-learn untuk persiapan data dan pandas.

3.4 Kerangka Model

Kerangka model memadukan pra-proses teks, embedding FastText, dan LSTM untuk regresi skor. Pada gambar 3.2 ditunjukkan kerangka model penelitian.



Gambar 3. 3 Kerangka Model Penelitian

Pada gambar 3.2 dapat. Kita lihat bahwa data yang dijadikan input (soal + jawaban) akan diproses terlebih dahulu untuk pembersihan dan normalisasi teks menggunakan casefolding dan tokenizer. Data teks yang sudah diubah menjadi

token dijadikan sebagai input oleh embeddings fasttext. Hasil dari embeddings fasttext berupa vektor angka dengan panjang 300 untuk setiap token. Sehingga hasil tersebut dijadikan sebagai input algoritma LSTM dengan fitur sebesar 300 dan sequence sebanyak jumlah kata dari kalimat terpanjang. Hasil pelatihan model algoritma LSTM akan dievaluasi performa kinerjanya menggunakan metrik MAPE.

BAB IV

HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil penelitian beserta pembahasannya yang disusun berdasarkan tiga aspek utama. Subbab 4.1 memaparkan proses dan hasil hyperparameter tuning beserta pembahasannya. Subbab 4.2 memaparkan hasil pelatihan dan pengujian model LSTM dan Bi-LSTM beserta pembahasannya. Subbab 4.3 memaparkan hasil pengujian berdasarkan karakteristik data beserta pembahasannya. Seluruh evaluasi performa model pada bab ini menggunakan metrik Mean Absolute Percentage Error (MAPE) sesuai yang dijelaskan pada Bab III.

4.1 Hyperparameter Tuning

Tahap pertama dalam penelitian ini adalah menentukan kombinasi hyperparameter optimal untuk model LSTM dan Bi-LSTM. Pengujian dilakukan secara sistematis dengan memvariasikan satu hyperparameter pada satu waktu (one-factor-at-a-time) sambil mempertahankan nilai default untuk hyperparameter lainnya. Empat kelompok hyperparameter yang diuji meliputi jumlah unit hidden (32, 64, 128, 256, 512), learning rate (0,0001; 0,0005; 0,001; 0,005; 0,01), batch size (16, 32, 64, 128), dan dropout rate (0,1; 0,2; 0,3; 0,4; 0,5). Setiap skenario dijalankan sebanyak 5 kali dengan data yang sama untuk mengurangi pengaruh inisialisasi bobot acak, kemudian dirata-ratakan hasilnya. Total terdapat 38 skenario pengujian yang mencakup 19 skenario untuk LSTM dan 19 skenario untuk Bi-LSTM.

Tabel 4.1 menyajikan ringkasan seluruh hasil pengujian hyperparameter tuning. Metrik evaluasi utama yang digunakan adalah Mean Absolute Percentage Error (MAPE) karena interpretasinya yang mudah dipahami sebagai persentase kesalahan rata-rata. Selain itu, standar deviasi MAPE turut dicantumkan untuk menggambarkan konsistensi model antar percobaan.

Tabel 4. 1 Hasil Pengujian Hyperparameter Tuning

No	Nama Skenario	Hidden	DR	LR	BS	Bidir.	MAPE (%)	Std. Dev.	Epoch
1	Unit Hidden 32	32	0.3	0.001	32	Tidak	9.5900	1.4517	83.6
2	Unit Hidden 64	64	0.3	0.001	32	Tidak	9.5898	2.4482	71.6
3	Unit Hidden 128	128	0.3	0.001	32	Tidak	9.2018	2.5354	67.4
4	Unit Hidden 256	256	0.3	0.001	32	Tidak	7.7596	2.4181	83.4
5	Unit Hidden 512	512	0.3	0.001	32	Tidak	11.0652	2.1132	36.0
6	Learning Rate 0.0001	128	0.3	0.0001	32	Tidak	9.1694	1.1479	100.0
7	Learning Rate 0.0005	128	0.3	0.0005	32	Tidak	7.1481	0.2511	90.8
8	Learning Rate 0.001	128	0.3	0.001	32	Tidak	8.2059	2.1204	82.4
9	Learning Rate 0.005	128	0.3	0.005	32	Tidak	12.0027	0.0111	22.4
10	Learning Rate 0.01	128	0.3	0.01	32	Tidak	11.6417	0.7759	36.0
11	Batch Size 16	128	0.3	0.001	16	Tidak	8.1527	2.2379	80.8
12	Batch Size 32	128	0.3	0.001	32	Tidak	9.3628	2.4679	68.6
13	Batch Size 64	128	0.3	0.001	64	Tidak	9.5363	2.2846	71.0
14	Batch Size 128	128	0.3	0.001	128	Tidak	11.1808	0.8397	81.6
15	Dropout Rate 0.1	128	0.1	0.001	32	Tidak	8.3584	2.1073	81.6
16	Dropout Rate 0.2	128	0.2	0.001	32	Tidak	10.0677	2.6765	52.0
17	Dropout Rate 0.3	128	0.3	0.001	32	Tidak	9.1108	2.6845	66.2
18	Dropout	128	0.4	0.001	32	Tidak	10.594	1.9292	46.2

No	Nama Skenario	Hidden	DR	LR	BS	Bidir.	MAPE (%)	Std. Dev.	Epoch
	Rate 0.4						1		
19	Dropout Rate 0.5	128	0.5	0.001	32	Tidak	8.6553	1.8960	84.2
20	Unit Hidden 32	32	0.3	0.001	32	Ya	7.9571	0.5190	90.6
21	Unit Hidden 64	64	0.3	0.001	32	Ya	7.5091	0.1554	95.0
22	Unit Hidden 128	128	0.3	0.001	32	Ya	7.4763	0.3037	84.6
23	Unit Hidden 256	256	0.3	0.001	32	Ya	8.4962	2.0198	79.6
24	Unit Hidden 512	512	0.3	0.001	32	Ya	11.985 7	0.0241	27.8
25	Learning Rate 0.0001	128	0.3	0.0001	32	Ya	7.7314	0.1695	100.0
26	Learning Rate 0.0005	128	0.3	0.0005	32	Ya	7.3432	0.1581	95.6
27	Learning Rate 0.001	128	0.3	0.001	32	Ya	7.1063	0.2629	89.4
28	Learning Rate 0.005	128	0.3	0.005	32	Ya	10.140 1	0.4131	81.2
29	Learning Rate 0.01	128	0.3	0.01	32	Ya	10.730 3	0.4592	49.6
30	Batch Size 16	128	0.3	0.001	16	Ya	6.8872	0.5175	89.2
31	Batch Size 32	128	0.3	0.001	32	Ya	7.4710	0.2093	84.8
32	Batch Size 64	128	0.3	0.001	64	Ya	10.157 4	2.5731	51.8
33	Batch Size 128	128	0.3	0.001	128	Ya	11.094 3	1.2904	62.6
34	Dropout Rate 0.1	128	0.1	0.001	32	Ya	7.3528	0.3200	87.8
35	Dropout Rate 0.2	128	0.2	0.001	32	Ya	8.2820	2.0379	83.2
36	Dropout Rate 0.3	128	0.3	0.001	32	Ya	7.5781	0.5796	88.6
37	Dropout Rate 0.4	128	0.4	0.001	32	Ya	7.4788	0.5263	95.2
38	Dropout Rate 0.5	128	0.5	0.001	32	Ya	8.2237	2.1570	78.2

Berdasarkan Tabel 4.1, pengujian jumlah unit hidden pada model LSTM menunjukkan bahwa konfigurasi dengan 256 unit menghasilkan MAPE terbaik sebesar 7,7596%, sedangkan konfigurasi dengan 512 unit menghasilkan performa terburuk (11,0652%) akibat *overfitting* yang ditandai dengan jumlah epoch yang lebih sedikit. Pada model Bi-LSTM, unit hidden 128 menghasilkan MAPE terbaik sebesar 7,4763%. Pengujian learning rate pada LSTM menunjukkan nilai 0,0005 memberikan hasil terbaik (MAPE 7,1481%), sementara pada Bi-LSTM nilai 0,001 menghasilkan MAPE terbaik 7,1063%. Learning rate yang terlalu besar (0,005 dan 0,01) mengakibatkan konvergensi prematur dengan epoch yang sangat sedikit (22–36 epoch).

Pengujian batch size menunjukkan batch size 16 menghasilkan performa terbaik pada kedua model: MAPE 8,1527% untuk LSTM dan 6,8872% untuk Bi-LSTM. Batch size yang lebih besar (128) menghasilkan performa lebih buruk karena estimasi gradien yang kasar. Pada pengujian dropout rate, nilai 0,1 memberikan MAPE terbaik pada LSTM (8,3584%) dan Bi-LSTM (7,3528%). Dropout rate yang terlalu tinggi (0,4–0,5) cenderung meningkatkan error karena regularisasi yang berlebihan.

Berdasarkan analisis menyeluruh terhadap seluruh skenario pengujian, konfigurasi hyperparameter terpilih untuk pelatihan model final ditentukan dengan mempertimbangkan keseimbangan antara performa, efisiensi komputasi, dan kemampuan generalisasi. Tabel 4.2 menampilkan konfigurasi hyperparameter yang dipilih beserta justifikasinya.

Tabel 4. 2 Konfigurasi Hyperparameter Terpilih

Hyperparameter	Nilai LSTM	Nilai Bi-LSTM	Justifikasi
Jumlah Unit Hidden	128	128	Keseimbangan kapasitas dan efisiensi
Learning Rate	0,001	0,001	Konvergensi stabil, epoch memadai
Batch Size	32	32	Kompromi antara kecepatan dan akurasi gradien
Dropout Rate	0,3	0,3	Regularisasi moderat, mencegah overfitting
Optimizer	Adam	Adam	Adaptif, efektif untuk data teks
Loss Function	MSE	MSE	Sesuai untuk regresi skor kontinu
Early Stopping	Patience=10	Patience=10	Mencegah overfitting dan mempersingkat waktu

Konfigurasi unit hidden sebesar 128 dipilih meskipun pengujian menunjukkan unit 256 memberikan MAPE sedikit lebih baik pada LSTM, karena mempertimbangkan efisiensi komputasi dan risiko overfitting pada dataset yang relatif kecil (500 sampel). Penentuan konfigurasi ini mengikuti prinsip Occam's Razor dalam machine learning bahwa model yang lebih sederhana dengan performa yang memadai lebih disukai daripada model yang kompleks.

4.2 Hasil Pelatihan dan Evaluasi Model

Setelah konfigurasi hyperparameter ditetapkan, model LSTM dan Bi-LSTM dilatih menggunakan seluruh data latih (400 sampel) dengan konfigurasi tersebut. Proses pelatihan dipantau melalui nilai loss pada data latih dan data validasi setiap epoch. Mekanisme early stopping dengan patience=10 diterapkan untuk menghentikan pelatihan ketika tidak ada perbaikan pada validation loss selama 10 epoch berturut-turut.

4.2.1 Model LSTM

Model LSTM menyelesaikan pelatihan pada epoch ke-100 dengan total parameter yang dapat dilatih sebanyak 227,969 parameter. Waktu pelatihan yang

dibutuhkan adalah 407.85 detik (~6.8 menit). Tabel 4.3 menyajikan data riwayat loss LSTM yang diambil pada epoch-epoch representatif untuk memperlihatkan pola konvergensi model.

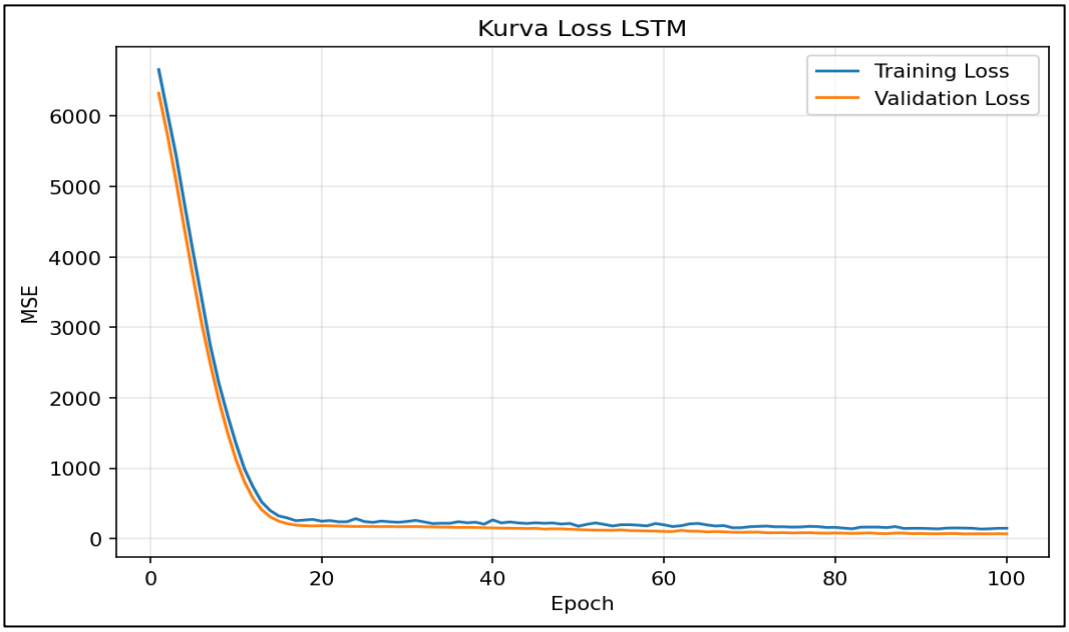
Tabel 4. 3 Riwayat Loss Model LSTM (Epoch Representatif)

Epoch	Training Loss (MSE)	Validation Loss (MSE)	Keterangan
1	6658.9395	6322.6816	Nilai awal, loss sangat tinggi
5	4073.4951	3716.0266	Penurunan cepat
10	1355.7887	1125.0801	Penurunan cepat
20	247.2319	183.6814	Stabilisasi awal
30	243.3512	169.9179	Stabilisasi awal
40	265.1287	151.3165	Konvergensi gradual
50	176.2890	128.2542	Konvergensi gradual
60	195.7758	102.1644	Konvergensi gradual
70	168.6313	91.2177	Konvergensi gradual
80	159.4785	79.6004	Konvergensi gradual
90	146.1756	72.4130	Konvergensi gradual
100	146.9607	67.4122	Epoch akhir (final)

Berdasarkan Tabel 4.3, model LSTM menunjukkan pola konvergensi yang baik. Pada epoch pertama, training loss sebesar 6,658.94 dan validation loss sebesar 6,322.68 yang merupakan nilai awal yang sangat tinggi. Penurunan loss terjadi sangat cepat pada epoch 1–10 karena model belajar pola dasar dari data. Pada epoch ke-100 (epoch akhir), training loss mencapai 146.9607 dan validation loss mencapai 67.4122, masing-masing mengalami penurunan sebesar 97.8% dan 98.9% dari nilai awal. Fakta bahwa validation loss lebih rendah dari training loss pada epoch-epoch akhir mengindikasikan bahwa model tidak mengalami overfitting dan mampu menggeneralisasi dengan baik.

Gambar 4.1 menampilkan kurva loss LSTM secara visual. Kurva training loss (biru) dan validation loss (oranye) memiliki pola penurunan yang serupa dan

bergerak beriringan sepanjang proses pelatihan, yang merupakan indikator yang baik bahwa model belajar secara konsisten tanpa terjadi overfitting yang signifikan.



Gambar 4. 1 Kurva Loss Model LSTM

Setelah pelatihan selesai, model LSTM dievaluasi pada data uji (75 sampel). Tabel 4.4 menyajikan metrik evaluasi lengkap model LSTM.

Tabel 4. 4 Metrik Evaluasi Model LSTM pada Data Uji

Metrik	Nilai	Keterangan
MAPE (%)	7.3747%	Kategori: Sangat Baik (< 10%)
MAE	5.8484	Rata-rata selisih absolut skor
RMSE	7.4653	Akar rata-rata kuadrat selisih
Jumlah Data Uji	75	Sampel uji (75 dari 500 total)
Epoch Aktual	100	Epoch penghentian dengan early stopping
Waktu Pelatihan	407.85 detik	Sekitar 6.8 menit
Trainable Parameters	227,969	Jumlah parameter yang dapat dilatih

Model LSTM menghasilkan MAPE sebesar 7.3747% pada data uji, yang masuk dalam kategori "Sangat Baik" (MAPE < 10%). Nilai MAE sebesar 5.8484

menunjukkan bahwa rata-rata selisih absolut antara skor prediksi dan skor aktual adalah sekitar 5.85 poin. RMSE sebesar 7.4653 sedikit lebih tinggi dari MAE, mengindikasikan adanya beberapa kasus prediksi yang memiliki selisih cukup besar sehingga meningkatkan nilai kuadrat error.

Tabel 4. 5 Hasil Prediksi Model LSTM (20 Data Pertama dari 75)

No	Aktual	Prediksi	Selisih Abs.	APE (%)	Keterangan
1	95	89.12	5.88	6.1868	Sangat Akurat
2	85	88.08	3.08	3.6232	Hampir Sempurna
3	90	84.34	5.66	6.2916	Sangat Akurat
4	95	88.65	6.35	6.6794	Sangat Akurat
5	95	89.82	5.18	5.4521	Sangat Akurat
6	80	85.85	5.85	7.3106	Sangat Akurat
7	95	88.20	6.80	7.1532	Sangat Akurat
8	80	83.64	3.64	4.5497	Hampir Sempurna
9	85	86.41	1.41	1.6591	Hampir Sempurna
10	80	81.08	1.08	1.3470	Hampir Sempurna
11	80	83.06	3.06	3.8202	Hampir Sempurna
12	50	54.21	4.21	8.4251	Sangat Akurat
13	95	89.89	5.11	5.3826	Sangat Akurat
14	55	58.91	3.91	7.1123	Sangat Akurat
15	90	86.19	3.81	4.2373	Hampir Sempurna
16	90	72.58	17.42	19.3563	Kurang Akurat
17	85	84.96	0.04	0.0493	Hampir Sempurna
18	95	87.92	7.08	7.4486	Sangat Akurat
19	95	89.65	5.35	5.6316	Sangat Akurat
20	80	76.36	3.64	4.5450	Hampir Sempurna

Tabel 4.5 memperlihatkan bahwa sebagian besar prediksi model LSTM berada dalam kisaran yang dekat dengan nilai aktual. Mayoritas prediksi dikategorikan sebagai "Hampir Sempurna" (APE < 5%) dan "Sangat Akurat" (APE 5%–10%). Terdapat satu kasus pada data ke-16 dengan APE sebesar 19,3563% (skor aktual 90, prediksi 72,58) yang dikategorikan "Kurang Akurat",

kemungkinan disebabkan oleh karakteristik jawaban yang berbeda secara signifikan dari pola umum dalam data latih.

Tabel 4. 6 Distribusi Error Model LSTM

Rentang APE	Jumlah Data	Persentase (%)	Keterangan
< 5%	28	37.33%	Hampir Sempurna
5% - 10%	31	41.33%	Sangat Akurat
10% - 20%	13	17.33%	Cukup Akurat
> 20%	3	4.00%	Kurang Akurat
Total < 10%	59	78.67%	Gabungan Sangat Baik

Berdasarkan distribusi error pada Tabel 4.6, model LSTM menunjukkan bahwa 37,33% prediksi memiliki APE di bawah 5% (kategori "Hampir Sempurna") dan 41,33% prediksi memiliki APE antara 5%–10% (kategori "Sangat Akurat"). Secara keseluruhan, 78,67% dari total 75 data uji memiliki APE di bawah 10%, yang menandakan model LSTM dapat memberikan prediksi yang andal untuk mayoritas sampel. Hanya 4,00% prediksi (3 data) yang memiliki APE lebih dari 20%, yang merupakan kasus-kasus ekstrem yang kemungkinan melibatkan jawaban dengan pola yang tidak terwakili dalam data latih.

4.2.2 Model Bi-LSTM

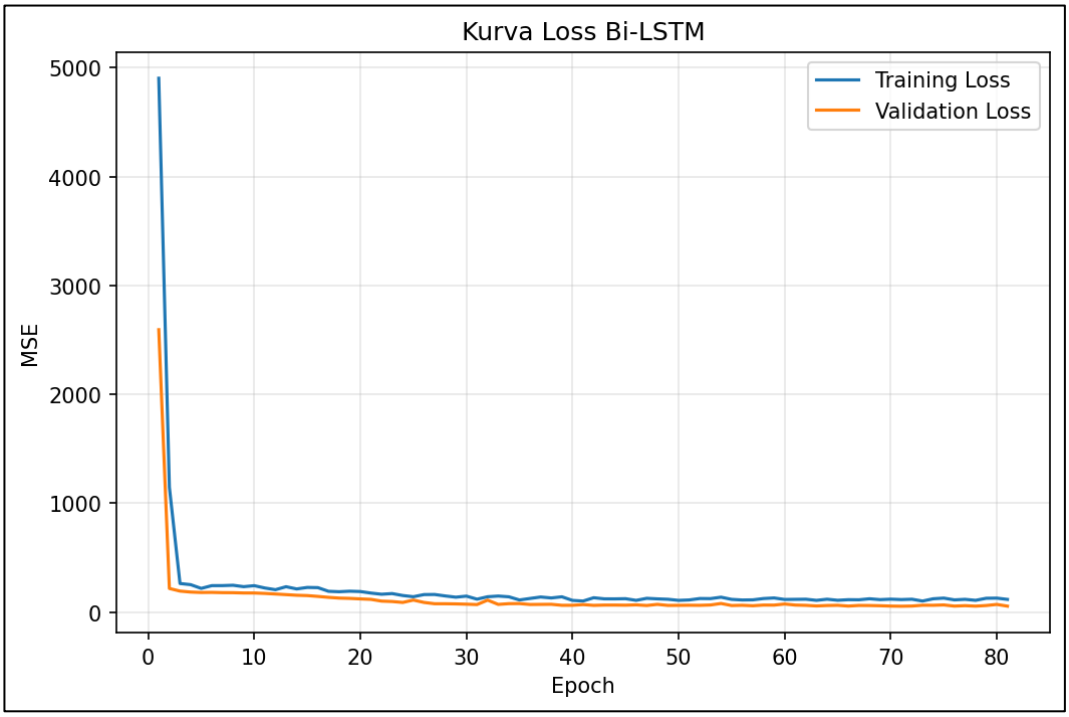
Model Bi-LSTM menyelesaikan pelatihan pada epoch ke-81 dengan total parameter yang dapat dilatih sebanyak 455,809 parameter sekitar dua kali lipat parameter LSTM (227,969) karena arsitektur bidirectional memproses urutan data secara maju (forward) dan mundur (backward) secara paralel. Waktu pelatihan yang dibutuhkan adalah 750.55 detik (~12.5 menit), lebih lama dibandingkan LSTM (407.85 detik) sesuai dengan kompleksitas arsitektur yang lebih tinggi.

Tabel 4. 7 Riwayat Loss Model Bi-LSTM (Epoch Representatif)

Epoch	Training Loss (MSE)	Validation Loss (MSE)	Keterangan
1	4902.8726	2593.5789	Nilai awal, loss sangat tinggi
5	217.6028	180.0949	Penurunan cepat
10	242.7104	175.3519	Penurunan cepat
20	188.8569	121.0925	Stabilisasi awal
30	146.6476	72.5751	Stabilisasi awal
40	107.5099	62.6306	Konvergensi gradual
50	107.7081	62.0781	Konvergensi gradual
60	115.9983	74.1404	Konvergensi gradual
70	118.7981	55.1544	Konvergensi gradual
81	116.6906	54.6057	Epoch akhir (early stop)

Pada epoch pertama, Bi-LSTM memiliki training loss 4,902.87 dan validation loss 2,593.58. Pada epoch akhir (ke-81), training loss turun menjadi 116.6906 dan validation loss menjadi 54.6057, masing-masing mengalami penurunan sebesar 97.6% dan 97.9% dari nilai awal. Model Bi-LSTM berhenti lebih awal (epoch 81) dibandingkan LSTM (epoch 100) karena mekanisme early stopping mengindikasikan bahwa model telah mencapai konvergensi optimal pada epoch tersebut.

Gambar 4.2 menampilkan kurva loss Bi-LSTM. Pola penurunan kurva Bi-LSTM mirip dengan LSTM namun memiliki pola awal yang berbeda validation loss Bi-LSTM pada epoch pertama (2.593,58) jauh lebih rendah dari training loss (4.902,87), yang menunjukkan model awalnya memprediksi data validasi dengan lebih baik sebelum bobot disesuaikan sepenuhnya. Kondisi ini normal pada model bidirectional karena konteks informasi yang lebih kaya dari arah maju dan mundur.



Gambar 4. 2 Kurva Loss Model Bi-LSTM

Tabel 4. 8 Metrik Evaluasi Model Bi-LSTM pada Data Uji

Metrik	Nilai	Keterangan
MAPE (%)	7.2696%	Kategori: Sangat Baik (< 10%)
MAE	5.5958	Rata-rata selisih absolut skor
RMSE	7.4171	Akar rata-rata kuadrat selisih
Jumlah Data Uji	75	Sampel uji (75 dari 500 total)
Epoch Aktual	81	Epoch penghentian dengan early stopping
Waktu Pelatihan	750.55 detik	Sekitar 12.5 menit
Trainable Parameters	455,809	Jumlah parameter yang dapat dilatih

Model Bi-LSTM menghasilkan MAPE sebesar 7.2696% pada data uji, juga masuk dalam kategori "Sangat Baik". Nilai MAE sebesar 5.5958 dan RMSE sebesar 7.4171 keduanya lebih rendah dibandingkan LSTM (MAE: 5.8484, RMSE: 7.4653), menunjukkan bahwa Bi-LSTM memiliki presisi prediksi yang sedikit lebih baik secara keseluruhan.

Tabel 4. 9 Hasil Prediksi Model Bi-LSTM (20 Data Pertama dari 75)

No	Aktual	Prediksi	Selisih Abs.	APE (%)	Keterangan
1	95	90.66	4.34	4.5655	Hampir Sempurna
2	85	88.04	3.04	3.5743	Hampir Sempurna
3	90	87.45	2.55	2.8374	Hampir Sempurna
4	95	90.72	4.28	4.5060	Hampir Sempurna
5	95	91.76	3.24	3.4097	Hampir Sempurna
6	80	86.70	6.70	8.3733	Sangat Akurat
7	95	90.81	4.19	4.4156	Hampir Sempurna
8	80	86.91	6.91	8.6416	Sangat Akurat
9	85	88.06	3.06	3.6006	Hampir Sempurna
10	80	85.98	5.98	7.4764	Sangat Akurat
11	80	83.66	3.66	4.5810	Hampir Sempurna
12	50	50.68	0.68	1.3580	Hampir Sempurna
13	95	91.54	3.46	3.6429	Hampir Sempurna
14	55	45.93	9.07	16.4933	Kurang Akurat
15	90	87.99	2.01	2.2368	Hampir Sempurna
16	90	80.38	9.62	10.6842	Cukup Akurat
17	85	87.75	2.75	3.2383	Hampir Sempurna
18	95	89.72	5.28	5.5594	Sangat Akurat
19	95	91.46	3.54	3.7223	Hampir Sempurna
20	80	81.95	1.95	2.4348	Hampir Sempurna

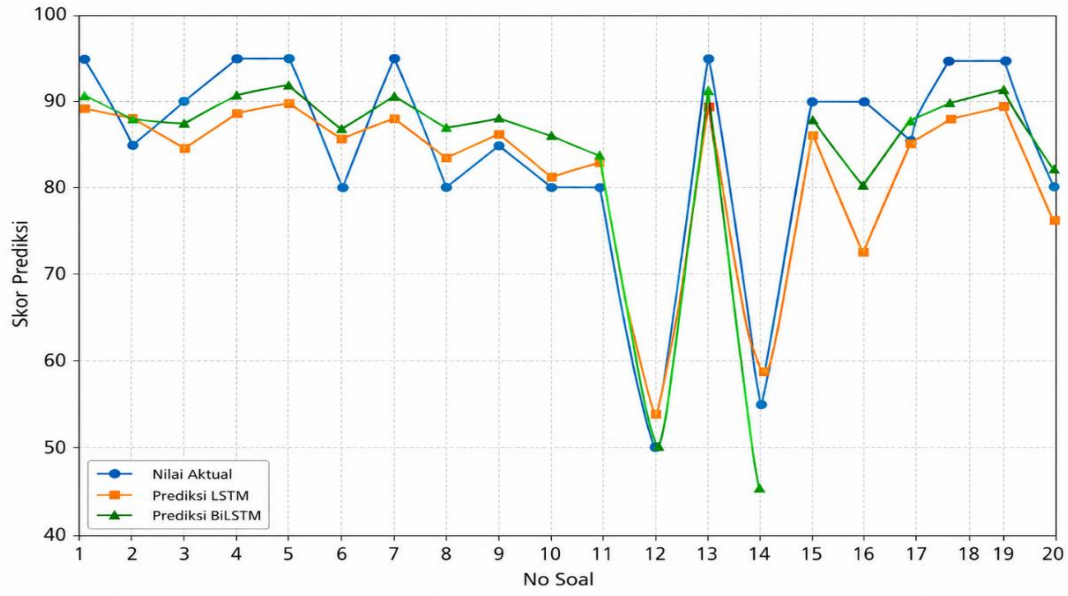
Tabel 4.9 menunjukkan bahwa Bi-LSTM menghasilkan lebih banyak prediksi dengan kategori "Hampir Sempurna" (APE < 5%) dibandingkan LSTM. Dari 20 data yang ditampilkan, mayoritas memiliki APE di bawah 5%, kecuali pada data ke-6 (APE 8,37%), ke-8 (APE 8,64%), ke-10 (APE 7,48%), dan data ke-16 (APE 10,68%) serta ke-14 (APE 16,49%). Kemampuan Bi-LSTM membaca konteks dua arah (forward dan backward) memungkinkannya menangkap fitur semantik yang lebih kaya dari representasi FastText.

4.2.3 Perbandingan Model LSTM dan Bi-LSTM

Setelah mengevaluasi kedua model secara terpisah, perbandingan komprehensif dilakukan untuk menentukan model yang lebih unggul dalam konteks penilaian esai otomatis. Tabel 4.10 menyajikan perbandingan metrik evaluasi dan karakteristik komputasi kedua model dan gambar 4. 3 ditunjukkan perbandingan hasil prediksi LSTM dan Bi LSTM dengan nilai aktualnya.

Tabel 4. 10 Perbandingan Komprehensif Model LSTM dan Bi-LSTM

Aspek Perbandingan	LSTM	Bi-LSTM	Keunggulan
MAPE (%)	7.3747%	7.2696%	Bi-LSTM
MAE	5.8484	5.5958	Bi-LSTM
RMSE	7.4653	7.4171	Bi-LSTM
Epoch Aktual	100	81	Bi-LSTM (lebih cepat konvergen)
Waktu Pelatihan (detik)	407.85	750.55	LSTM (lebih cepat)
Trainable Parameters	227,969	455,809	LSTM (lebih efisien)
Training Loss Akhir (MSE)	146.9607	116.6906	Bi-LSTM
Validation Loss Akhir (MSE)	67.4122	54.6057	Bi-LSTM



Gambar 4. 3 Hasil Prediksi Model LSTM dengan Bi LSTM Berdasarkan Nilai Aktual

Tabel 4.10 menunjukkan bahwa Bi-LSTM unggul dalam metrik akurasi prediksi (MAPE, MAE, RMSE) dibandingkan LSTM. Selisih MAPE antara keduanya adalah 0.1051 persentase poin (7.3747% vs 7.2696%), selisih MAE sebesar 0.2525 poin skor, dan selisih RMSE sebesar 0.0482. Meskipun perbedaan ini tergolong kecil, secara konsisten Bi-LSTM memberikan prediksi yang lebih presisi pada semua metrik evaluasi yang digunakan.

Dari sisi efisiensi komputasi, LSTM lebih unggul dengan waktu pelatihan 407.85 detik dibandingkan Bi-LSTM yang membutuhkan 750.55 detik (selisih 343 detik atau 84.0% lebih lama). Jumlah parameter Bi-LSTM (455,809) juga hampir dua kali lipat LSTM (227,969) karena setiap layer bidirectional memiliki dua set bobot (forward dan backward). Pertimbangan ini penting dalam konteks deployment sistem penilaian esai pada lingkungan dengan sumber daya komputasi terbatas.

Tabel 4. 11 Distribusi Error Komparatif LSTM vs Bi-LSTM

Rentang APE	LSTM (n)	LSTM (%)	BiLSTM (n)	BiLSTM (%)	Selisih (%)
< 5%	28	37.33%	39	52.00%	+14.67%
5% - 10%	31	41.33%	20	26.67%	-14.67%
10% - 20%	13	17.33%	10	13.33%	-4.00%
> 20%	3	4.00%	6	8.00%	+4.00%
Total < 10%	59	78.67%	59	78.67%	0.00%

Berdasarkan Tabel 4.11, Bi-LSTM unggul dalam proporsi prediksi "Hampir Sempurna" (APE < 5%) dengan 52,00% berbanding 37,33% pada LSTM, meningkat sebesar 14,67 persentase poin. Namun, Bi-LSTM memiliki lebih banyak prediksi dengan APE > 20% (8,00% vs 4,00%), mengindikasikan bahwa meskipun Bi-LSTM lebih presisi secara rata-rata, model ini memiliki beberapa

kasus prediksi yang lebih meleset. Proporsi prediksi dengan APE < 10% (kategori gabungan "Sangat Baik") sama persis antara kedua model (78,67%), menunjukkan bahwa keduanya memiliki kemampuan yang setara dalam menghasilkan prediksi berkualitas tinggi pada mayoritas data uji, perbandingan pengujian .

4.3 Pengujian Berdasarkan Karakteristik Data

Untuk memahami perilaku model secara lebih mendalam, pengujian lanjutan dilakukan dengan membagi data berdasarkan tiga dimensi karakteristik: (1) panjang kalimat jawaban, (2) kelompok skor berdasarkan Kriteria Ketuntasan Minimum (KKM), dan (3) skema pembagian data uji. Pengujian ini bertujuan mengidentifikasi kondisi di mana model bekerja optimal dan kondisi di mana model mengalami kesulitan, sehingga dapat memberikan panduan yang lebih baik untuk pengembangan sistem ke depannya.

4.3.1 Pengujian Berdasarkan Panjang Kalimat

Jawaban esai siswa memiliki variasi panjang yang cukup besar, dari minimal 2 kata hingga maksimal 53 kata. Untuk analisis ini, data dibagi menjadi dua kelompok: kalimat pendek (< 26 kata) dan kalimat panjang (≥ 26 kata) berdasarkan median panjang jawaban. Tabel 4.12 menyajikan statistik deskriptif kedua kelompok tersebut.

Tabel 4. 12 Statistik Deskriptif Kelompok Panjang Kalimat

Kelompok	N	Min (kata)	Maks (kata)	Rata-rata Panjang	Rata-rata Skor	Keterangan
Kalimat Pendek (< 26 kata)	395	2	25	15.78	82.02	Mayoritas data
Kalimat Panjang (>= 26 kata)	105	26	53	30.62	88.18	Minoritas, skor lebih tinggi
Keseluruhan	500	2	53	18.89	83.31	Seluruh dataset

Dari Tabel 4.12, kalimat pendek mendominasi dataset dengan 395 dari 500 data (79,00%), sementara kalimat panjang berjumlah 105 data (21,00%). Rata-rata skor kalimat panjang (88,18) lebih tinggi dari kalimat pendek (82,02), mengindikasikan korelasi positif antara panjang jawaban dan kualitas/skor esai. Hal ini sejalan dengan harapan bahwa jawaban yang lebih elaboratif dan panjang umumnya memperoleh skor lebih tinggi.

Tabel 4. 13 Hasil Pengujian MAPE Berdasarkan Panjang Kalimat

Subkelompok	N	MAPE LSTM (%)	MAPE Bi-LSTM (%)	Selisih MAPE	Keunggulan
Kalimat Pendek	395	8.6332%	7.5136%	1.1196%	Bi-LSTM
Kalimat Panjang	105	5.1208%	4.8336%	0.2872%	Bi-LSTM
Keseluruhan	500	7.8956%	6.9508%	0.9448%	Bi-LSTM

Tabel 4.13 menunjukkan pola yang konsisten pada kedua model: prediksi pada kalimat panjang jauh lebih akurat dibandingkan kalimat pendek. MAPE LSTM pada kalimat panjang (5,1208%) lebih rendah 3,5124 persentase poin dibandingkan kalimat pendek (8,6332%). Demikian pula Bi-LSTM pada kalimat panjang (4,8336%) vs kalimat pendek (7,5136%). Temuan ini dapat dijelaskan

melalui karakteristik representasi FastText: jawaban yang lebih panjang menghasilkan vektor rata-rata yang lebih stabil dan representatif karena rata-rata dari lebih banyak vektor kata cenderung menangkap semantik keseluruhan kalimat dengan lebih baik. Sebaliknya, jawaban pendek (2–5 kata) menghasilkan vektor yang kurang informatif dan lebih rentan terhadap pengaruh kata-kata individual yang tidak relevan.

4.3.2 Pengujian Berdasarkan Kelompok Skor

Data dibagi berdasarkan posisi skor terhadap KKM (Kriteria Ketuntasan Minimum) sebesar 75. Tiga kelompok dibentuk: di atas KKM (skor > 75), tepat KKM (skor = 75), dan di bawah KKM (skor < 75). Distribusi data menunjukkan ketidakseimbangan yang cukup signifikan: kelompok di atas KKM mendominasi dengan 400 data (80,00%), tepat KKM 44 data (8,80%), dan di bawah KKM 56 data (11,20%). Tabel 4.14 menyajikan hasil pengujian MAPE per kelompok skor.

Tabel 4. 14 Hasil Pengujian MAPE Berdasarkan Kelompok Skor KKM

Subkelompok	N	MAPE LSTM (%)	MAPE Bi-LSTM (%)	Selisih MAPE	Keunggulan
Di atas KKM (> 75)	400	5.8912%	4.9292%	0.9619%	Bi-LSTM
Tepat KKM (= 75)	44	7.9058%	10.3232%	2.4173%	LSTM
Di bawah KKM (< 75)	56	22.2048%	18.7407%	3.4640%	Bi-LSTM
Keseluruhan	500	7.8956%	6.9508%	0.9448%	Bi-LSTM

Tabel 4.14 mengungkap temuan kritis: model mengalami kesulitan signifikan pada kelompok skor di bawah KKM. MAPE LSTM pada kelompok di bawah KKM (22,2048%) hampir 4 kali lebih tinggi dibandingkan kelompok di atas

KKM (5,8912%). Pola serupa terjadi pada Bi-LSTM (18,7407% vs 4,9292%). Fenomena ini disebabkan oleh ketidakseimbangan kelas dalam data latih: karena 80% data berada di atas KKM, model lebih terbiasa dengan pola jawaban berkualitas tinggi dan kurang terekspos pada pola jawaban dengan skor rendah.

Temuan menarik lainnya adalah pada kelompok tepat KKM (skor = 75): model LSTM (MAPE 7,9058%) ternyata lebih baik dari Bi-LSTM (MAPE 10,3232%) dengan selisih 2,4173 persentase poin. Kondisi ini kemungkinan disebabkan oleh ambiguitas fitur pada skor batas jawaban dengan skor tepat 75 berada di perbatasan antara "lulus" dan "tidak lulus", sehingga Bi-LSTM yang lebih sensitif terhadap konteks dua arah justru lebih mudah terpengaruh oleh variasi kecil dalam pola teks, sedangkan LSTM dengan konteks satu arah lebih stabil dalam kondisi ini.

Dari perspektif keseluruhan dataset, kedua model menunjukkan MAPE yang sama (7,8956% LSTM dan 6,9508% Bi-LSTM) karena dominasi data di atas KKM. Temuan ini menjadi rekomendasi penting bahwa untuk implementasi yang lebih adil, teknik penanganan ketidakseimbangan data seperti oversampling atau penambahan bobot pada kelas minoritas perlu dipertimbangkan dalam pengembangan model ke depan.

4.3.3 Pengujian Berdasarkan Skema Pembagian Data

Pengujian dengan skema pembagian data yang berbeda dilakukan untuk mengukur kemampuan generalisasi model dalam skenario penggunaan yang realistis. Tiga skema digunakan: (1) Berdasarkan Siswa data uji terdiri dari

jawaban siswa-siswa yang tidak termasuk dalam data latih (75 data dari 15 siswa baru); (2) Berdasarkan Soal data uji berisi semua jawaban untuk soal-soal yang tidak termasuk dalam data latih (100 data dari 1 soal); (3) Acak atau Random pembagian data secara acak tanpa mempertimbangkan siswa atau soal (75 data, skema standar yang digunakan pada evaluasi utama). Tabel 4.15 menyajikan hasilnya.

Tabel 4. 15 Hasil Pengujian MAPE Berdasarkan Skema Pembagian Data

Skema	N Uji	MAPE LSTM (%)	MAPE Bi-LSTM (%)	Selisih MAPE	Keunggulan
Berdasarkan Siswa	75	6.1671%	9.3566%	3.1894%	LSTM
Berdasarkan Soal	100	16.0616%	12.0010%	4.0606%	Bi-LSTM
Acak / Random	75	7.2747%	7.6143%	0.3396%	LSTM

Hasil pada Tabel 4.15 menunjukkan variasi performa yang sangat signifikan antarskema. Skema berdasarkan soal menghasilkan MAPE paling tinggi pada kedua model (LSTM: 16,0616%, Bi-LSTM: 12,0010%), mengindikasikan bahwa model mengalami kesulitan besar saat harus menilai esai untuk soal yang sama sekali belum pernah dilihat dalam pelatihan. Ini wajar karena setiap soal memiliki domain kosakata dan pola jawaban yang unik model yang hanya dilatih dengan 4 soal tidak dapat menggeneralisasi dengan baik ke soal ke-5 yang berbeda secara semantik.

Skema berdasarkan siswa (LSTM: 6,1671%, Bi-LSTM: 9,3566%) dan skema acak (LSTM: 7,2747%, Bi-LSTM: 7,6143%) menunjukkan performa yang jauh lebih baik. Pada skema berdasarkan siswa, LSTM (6,1671%) unggul atas Bi-LSTM (9,3566%) dengan selisih 3,1894 persentase poin anomali menarik di mana LSTM lebih unggul dari Bi-LSTM. Hal ini menunjukkan bahwa kemampuan

LSTM dalam mempertahankan memori sekuensial jangka panjang menjadi keunggulan saat menghadapi siswa-siswa baru yang pola penulisannya mungkin berbeda dari siswa latih.

Skema acak menghasilkan MAPE yang relatif seimbang antara kedua model (LSTM: 7,2747%, Bi-LSTM: 7,6143%), dengan LSTM sedikit lebih unggul. Temuan ini mengonfirmasi bahwa hasil evaluasi utama (pembagian acak 80:15 latih:uji) merupakan skenario yang paling optimis karena data latih dan uji berasal dari distribusi yang sama, sementara skenario berdasarkan soal memberikan gambaran tantangan yang lebih realistis dalam deployment sistem AES.

Tabel 4. 16 Ringkasan Hasil Pengujian Karakteristik Data

Dimensi	Subkelompok	MAPE LSTM (%)	MAPE Bi-LSTM (%)	Model Unggul
Panjang Kalimat	Kalimat Pendek	8.6332%	7.5136%	Bi-LSTM
Panjang Kalimat	Kalimat Panjang	5.1208%	4.8336%	Bi-LSTM
Panjang Kalimat	Keseluruhan	7.8956%	6.9508%	Bi-LSTM
Kelompok Skor	Di atas KKM (> 75)	5.8912%	4.9292%	Bi-LSTM
Kelompok Skor	Tepat KKM (= 75)	7.9058%	10.3232%	LSTM
Kelompok Skor	Di bawah KKM (< 75)	22.2048%	18.7407%	Bi-LSTM
Kelompok Skor	Keseluruhan	7.8956%	6.9508%	Bi-LSTM
Skema Pembagian	Berdasarkan Siswa	6.1671%	9.3566%	LSTM
Skema Pembagian	Berdasarkan Soal	16.0616%	12.0010%	Bi-LSTM
Skema Pembagian	Acak / Random	7.2747%	7.6143%	LSTM

Tabel 4.16 merangkum seluruh hasil pengujian karakteristik data. Secara umum, Bi-LSTM lebih unggul pada sebagian besar kondisi pengujian. Namun, LSTM menunjukkan keunggulan pada dua kondisi spesifik yaitu kelompok tepat KKM dan skema pembagian berdasarkan siswa. Temuan ini memperkuat

kesimpulan bahwa pilihan model yang optimal bergantung pada konteks penggunaan sistem AES yang diinginkan.

4.4 Integrasi Islam dan Sains

Penelitian ini melalui penggabungan antara nilai-nilai Islam, ilmu pengetahuan, dan teknologi kecerdasan buatan dalam bidang pendidikan. Penelitian ini tidak hanya berfokus pada pengembangan model prediksi nilai jawaban esai, tetapi juga diarahkan untuk mendukung proses penilaian yang lebih objektif, konsisten, efisien, dan adil. Hal ini sejalan dengan nilai Islam tentang pentingnya menuntut ilmu, amanah, dan keadilan dalam mengambil keputusan, terutama dalam konteks penilaian hasil belajar peserta didik.

Pada bagian ini, dibahas refleksi filosofis dan teologis mengenai implementasi teknologi kecerdasan buatan (Artificial Intelligence) yang digunakan dalam penelitian ini. Integrasi antara sains modern (pemrosesan bahasa alami menggunakan FastText dan algoritma Long Short-Term Memory) dengan nilai-nilai universal Islam dianalisis melalui pendekatan dalil Al-Qur'an dan Hadis Nabi SAW yang relevan dengan konsep perekaman data, memori sekuensial, serta keadilan dalam penilaian.

4.4.1 Konsep Rekaman Sekuensial dan Arsitektur LSTM dalam QS. Al An'am (6): 59

Pemilihan algoritma Long Short-Term Memory (LSTM) dalam memprediksi nilai jawaban esai didasarkan pada kemampuannya untuk

memproses data teks yang bersifat sekuensial (berurutan) tanpa kehilangan konteks masa lalu (vanishing gradient problem). Secara epistemologi Islam, konsep pemeliharaan data secara runtut, sistematis, dan menyeluruh ini selaras dengan firman Allah SWT dalam QS. Al-An'am (6) ayat 59:

۞ وَعِنْدَهٗ مَفَاتِحُ الْغَيْبِ لَا يَعْلَمُهَآ اِلَّا هُوَۗ وَيَعْلَمُ مَا فِى الْبَرِّ وَالْبَحْرِۗ وَمَا تَسْقُطُ مِنْ وَّرَقَةٍ اِلَّا يَعْلَمُهَا وَلَا حَبَّةٍ فِيْ ظُلُمٰتِ الْاَرْضِ وَلَا رَطْبٍ وَّلَا يَابِسٍ اِلَّا فِيْ كِتٰبٍ مُّبِيْنٍ ۝٥٩

Artinya: "Dan pada sisi Allah-lah kunci-kunci semua yang ghaib; tidak ada yang mengetahuinya kecuali Dia sendiri, dan Dia mengetahui apa yang di daratan dan di lautan, dan tiada sehelai daun pun yang gugur melainkan Dia mengetahuinya (pula), dan tidak jatuh sebutir biji pun dalam kegelapan bumi dan tidak sesuatu yang basah atau yang kering, melainkan tertulis dalam kitab yang nyata (Lauh Mahfuzh)."

Secara komputasional, terdapat korelasi fungsional yang mendalam antara redaksi ayat tersebut dengan mekanisme kerja komponen pada algoritma LSTM, yang dijabarkan sebagai berikut:

1) Prinsip Data Input dan Observasi Universal: "Dia Allah mengetahui apa yang di daratan dan di lautan" merepresentasikan proses penerimaan input data dari berbagai ranah (domain). Dalam penelitian ini, sistem menerima input berupa beragam variasi teks jawaban esai yang dikonversi menjadi vektor numerik melalui metode FastText.
2) Prinsip Forget Gate dan Seleksi Informasi Spesifik: Merujuk pada pemrosesan kejadian-kejadian mikro yang sangat spesifik dan dinamis di alam semesta. Pada LSTM, hal ini diimplementasikan melalui forget gate, di mana algoritma secara cerdas memilah dan melupakan informasi yang

tidak relevan (noise / kata sambung), namun tetap mempertahankan kata kunci (keywords) penting yang ditulis oleh mahasiswa demi akurasi makna.

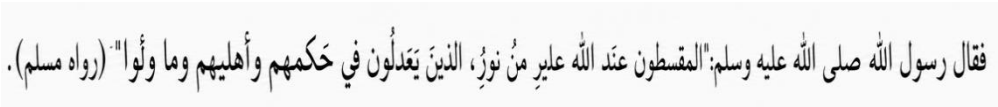
3) Prinsip Cell State dan Kitabum Mubin (Lauh Mahfuzh): Frasa "melainkan tertulis dalam kitab yang nyata (Lauh Mahfuzh)" menegaskan adanya sebuah sistem memori pusat yang bersifat permanen, runtut, dan tidak mengalami degradasi informasi seiring berjalannya waktu. Cell state pada LSTM mengadopsi prinsip ini dengan bertindak sebagai jalur memori jangka panjang (long-term memory loop). Ia membawa informasi tekstual dari awal kalimat hingga akhir paragraf esai agar konteks utuh dari jawaban tidak hilang saat proses prediksi nilai dilakukan.

4.4.2 Prinsip Keadilan dan Akurasi Penilaian Otomatis dalam Hadis Nabi Shallallahu ‘alaihi wa sallam

Tujuan akhir dari integrasi FastText dan LSTM dalam tesis ini adalah untuk membangun sistem penilaian yang memiliki tingkat akurasi tinggi dan objektif. Penilaian jawaban berbentuk esai oleh korektor manusia kerap kali dihadapkan pada tantangan subjektivitas, faktor kelelahan fisik (fatigue), dan inkonsistensi. Dalam perspektif Islam, memberikan penilaian atau keputusan akademik terhadap hasil belajar seseorang merupakan bagian dari amanah yang harus ditunaikan secara adil dan presisi.

Hal ini sejalan dengan hadis Rasulullah Shallallahu ‘alaihi wa sallam mengenai pentingnya profesionalisme, kompetensi, dan keadilan dalam

memberikan keputusan hukum atau penilaian, sebagaimana diriwayatkan oleh Abdullah bin Amr bin Al-Ash:


فقال رسول الله صلى الله عليه وسلم:"المقسطون عند الله عليرِ منُ نورُ، الذينَ يَعدلُون في حَكمهم وأهليهم وما ولُوا" (رواه مسلم).

‘Rasulullah Shallallahu ‘alaihi wa sallam bersabda: "Sesungguhnya orang-orang yang berlaku adil di sisi Allah berada di atas mimbar-mimbar yang terbuat dari cahaya. Mereka adalah orang-orang yang berlaku adil dalam menetapkan hukum/keputusan, adil terhadap keluarga mereka, dan adil terhadap apa saja yang diserahkan di bawah kekuasaan mereka." (HR. Muslim, No. 1827).

Implementasi hadis ini diwujudkan melalui peran kecerdasan buatan sebagai alat bantu (tool) untuk menegakkan nilai keadilan (computational justice). Sistem berbasis LSTM yang telah dilatih menggunakan representasi kata FastText akan memberikan penilaian berdasarkan parameter matematis yang objektif. Ketika model berhasil meminimalkan nilai error (seperti Mean Squared Error atau Mean Absolute Error), sistem tersebut secara langsung membantu institusi pendidikan dalam mendistribusikan hak nilai siswa secara adil. Penilaian otomatis ini menjamin bahwa setiap siswa diuji dengan standar yang sama, kapan pun esai mereka diperiksa, bebas dari bias manusiawi, yang merupakan representasi nyata dari pemanfaatan sains demi kemaslahatan umat (maslahah mursalah).

Penelitian ini menunjukkan adanya sinergi antara pengembangan teknologi Artificial Intelligence dengan nilai-nilai Islam. Secara teologis, arsitektur Long Short-Term Memory (LSTM) yang dipadukan dengan FastText merefleksikan konsep pencatatan data yang komprehensif, terstruktur, dan akurat sebagaimana tersirat dalam QS. Al-An'am (6): 59. Implementasi algoritma ini tidak hanya

berfungsi sebagai alat teknis, tetapi juga menjadi pengejawantahan prinsip keadilan (computational justice) dalam penilaian akademik, sejalan dengan anjuran Hadis Riwayat Muslim No. 1827 untuk berlaku adil dan objektif. Dengan meminimalkan bias subjektif serta inkonsistensi manusia, sistem ini menjadi bentuk ikhtiar ilmiah yang bernilai maslahah mursalah, yaitu pemanfaatan sains demi mewujudkan standarisasi pendidikan yang adil, presisi, dan bermanfaat bagi kemaslahatan umat.

BAB V

PENUTUP

5.1 Kesimpulan

Penelitian ini berhasil mengembangkan sistem Automated Essay Scoring (AES) berbasis deep learning yang mampu memprediksi skor jawaban esai siswa SMP Negeri 6 Singosari pada mata pelajaran Bahasa Inggris dengan akurasi tinggi. Adapun kesimpulan pada penelitian ini adalah sebagai berikut :

1. Penelitian ini membuktikan bahwa skor jawaban esai Bahasa Inggris siswa SMP dapat diprediksi secara otomatis menggunakan pendekatan deep learning berbasis representasi teks FastText. Inovasi utama berupa penggabungan teks soal dan jawaban dengan token khusus (<startanswer> dan <endanswer>) terbukti efektif memungkinkan model memahami relevansi jawaban terhadap konteks pertanyaan secara simultan. Model terbaik yang dihasilkan, yaitu Bi-LSTM dengan representasi FastText, mencapai MAPE sebesar 7.2696% pada data uji yang tergolong kategori "Sangat Baik" (MAPE < 10%) menurut interpretasi standar MAPE. Nilai ini menunjukkan bahwa rata-rata selisih antara skor prediksi dan skor aktual guru hanya sekitar 7.27% dari nilai skor aktual. Sebanyak 78,67% dari 75 data uji memiliki APE di bawah 10%, dan 52,00% bahkan di bawah 5%, membuktikan konsistensi dan keandalan sistem yang dibangun.
2. Pengujian berdasarkan karakteristik data mengungkapkan pola perilaku model yang konsisten. Berdasarkan panjang kalimat, sistem bekerja optimal pada jawaban panjang (≥ 26 kata) dengan MAPE Bi-LSTM 4.8336% dibandingkan

7.5136% pada jawaban pendek (< 26 kata). Berdasarkan kelompok skor KKM, sistem sangat andal pada jawaban di atas KKM (MAPE Bi-LSTM 4.9292%) namun menghadapi tantangan pada jawaban di bawah KKM (MAPE Bi-LSTM 18.7407%) akibat ketidakseimbangan distribusi kelas dalam dataset. Pada skema pembagian data berdasarkan soal, model menunjukkan penurunan performa yang signifikan (MAPE LSTM 16.0616%, Bi-LSTM 12.0010%), mengindikasikan tantangan generalisasi antar soal yang berbeda secara semantik.

3. Proses hyperparameter tuning yang dilakukan secara sistematis melalui 38 skenario pengujian berhasil mengidentifikasi konfigurasi optimal: 128 hidden units, learning rate 0,001, batch size 32, dan dropout rate 0,3. Model LSTM unidireksional dengan konfigurasi ini menghasilkan MAPE 7.3747% (kategori "Sangat Baik"), sementara model Bi-LSTM menghasilkan MAPE 7.2696%, lebih baik 0.1051 persentase poin dibandingkan LSTM. Perbandingan menyeluruh menunjukkan Bi-LSTM unggul pada seluruh metrik akurasi: MAE 5.5958 vs 5.8484 (LSTM), dan RMSE 7.4171 vs 7.4653 (LSTM). Namun trade-off antara Bi-LSTM dan LSTM perlu dipertimbangkan dalam implementasi: Bi-LSTM membutuhkan waktu pelatihan lebih lama (750.55 detik vs 407.85 detik, selisih 84.0% lebih lama) dan memiliki jumlah parameter yang lebih banyak (455,809 vs 227,969), namun secara konsisten menghasilkan akurasi yang lebih baik pada sebagian besar kondisi pengujian.

5.2 Saran

Berdasarkan hasil penelitian, keterbatasan yang ditemukan, dan potensi pengembangan yang teridentifikasi, sehingga didapat saran sebagai berikut :

1. Peningkatan ukuran dan keragaman dataset merupakan prioritas utama untuk pengembangan lanjutan. Hasil pengujian berdasarkan skema pembagian data menunjukkan bahwa model mengalami kesulitan signifikan pada soal yang tidak terlihat dalam pelatihan (MAPE hingga 16.06% untuk LSTM dan 12.00% untuk Bi-LSTM), mengindikasikan perlunya data latih yang lebih beragam secara semantik. Dataset yang lebih besar — idealnya ribuan hingga puluhan ribu pasangan soal-jawaban — akan memungkinkan model mempelajari pola yang lebih beragam dan meningkatkan kemampuan generalisasi. Ekspansi dataset untuk mencakup berbagai tingkat pendidikan (SD, SMP, SMA), berbagai mata pelajaran (tidak hanya Bahasa Inggris), dan berbagai sekolah dari latar belakang demografis yang berbeda akan menghasilkan sistem AES yang lebih universal.
2. Penanganan ketidakseimbangan distribusi kelas skor perlu diperhatikan dalam pengembangan model selanjutnya. Temuan penelitian menunjukkan bahwa model mengalami kesulitan signifikan pada jawaban dengan skor di bawah KKM (MAPE LSTM 22.20%, Bi-LSTM 18.74%) dibandingkan jawaban di atas KKM, yang disebabkan oleh dominasi data berkualitas tinggi (80% data berada di atas KKM) dalam dataset. Teknik-teknik penanganan ketidakseimbangan seperti oversampling (SMOTE), undersampling, atau pemberian bobot berbeda pada kelas minoritas saat pelatihan dapat

dieksplorasi untuk menghasilkan model yang lebih adil dan akurat di seluruh rentang skor. Hal ini penting agar sistem AES dapat memberikan penilaian yang reliabel bagi seluruh siswa, termasuk mereka yang mengalami kesulitan belajar.

3. Eksplorasi arsitektur berbasis transformer seperti BERT (Bidirectional Encoder Representations from Transformers), IndoBERT, atau model bahasa besar lainnya yang telah dilatih pada korpus Bahasa Indonesia dan Bahasa Inggris berpotensi menghasilkan representasi semantik yang lebih kaya dibandingkan FastText. Mekanisme self-attention dalam transformer memungkinkan model menangkap hubungan kontekstual jangka panjang antar token secara lebih efektif dibandingkan LSTM, yang dapat mengatasi keterbatasan model saat ini dalam menangani jawaban pendek dan jawaban di bawah KKM. Selain itu, pengembangan sistem penilaian multi-aspek yang memberikan skor terpisah untuk dimensi evaluasi seperti isi (content), tata bahasa (grammar), kosakata (vocabulary), dan koherensi (coherence) melalui pendekatan multi-task learning akan memberikan umpan balik yang jauh lebih kaya dan bermakna bagi siswa maupun guru.

DAFTAR PUSTAKA

Ainul Yaqin, M., & Abdul Charis Fauzan, Mk. (2025). *Teknologi Pengolahan Bahasa Alami: Konsep dan Algoritma*. www.betaaksara.my.id

Anđelković, J. (2022). Peer And Teacher Assessment Of Academic Essay Writing: Procedure And Correspondence. In *Journal of Teaching English for Specific and Academic Purposes* (Vol. 10, Issue 1, pp. 171–184). University of Nis. https://doi.org/10.22190/JTESAP2201171A

Arı, G. (2021). An Overview of Writing Rubrics in Doctoral Dissertations in Turkey. *International Journal of Progressive Education*, *17*(2), 385–405. https://doi.org/10.29329/ijpe.2021.332.24

Asli, N. F., Mohd Matore, M. E. E., & Md Yunus, M. (2024). Construct validity of primary trait writing rubrics based on assessment use argument (AUA) validation framework. *Heliyon*, *10*(22). https://doi.org/10.1016/j.heliyon.2024.e40053

Avcu, A. (2025). Employing a Hierarchical Rater Models for Automated Scoring: Scope Review on the Application in Educational Assessment. *Malaysian Online Journal of Educational Technology*, *13*(2), 37–47. https://doi.org/10.52380/mojet.2025.13.2.569

Baharudin, H., Maskor, Z. M., & Matore, M. E. E. M. (2022). The raters' differences in Arabic writing rubrics through the Many-Facet Rasch measurement model. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.988272

Balaha, H. M., & Saafan, M. M. (2021). Automatic exam correction framework (AECF) for the MCQS, essays, and equations matching. *IEEE Access*, *9*, 32368–32389. https://doi.org/10.1109/ACCESS.2021.3060940

Buditjahjanto, I. G. P. A., Idhom, M., Munoto, M., & Samani, M. (2022). An Automated Essay Scoring Based on Neural Networks to Predict and Classify Competence of Examinees in Community Academy. *TEM Journal*, *11*(4), 1694–1701. https://doi.org/10.18421/TEM114-34

Chamidah, N., Mega Santoni, M., Nurramdhani Irmanda, H., Astriratma, R., Ilmu Komputer, F., & Pembangunan Nasional Veteran Jakarta, U. (2022). *Penilaian Esai Pendek Otomatis Berdasarkan Similaritas Semantik dengan SBERT Semantic Similarity-Based Short Essay Auto Scoring using SBERT* (Vol. 21, Issue 4).

Dabbagh, A., & Safaei, A. (2019). Comparative Textbook Evaluation: Representation of Learning Objectives in Locally and Internationally Published ELT Textbooks. In *Issues in Language Teaching (ILT)* (Vol. 8, Issue 1).

Putri, F., & Zakir, S. (2023). Mengukur Keberhasilan Evaluasi Pembelajaran: Telaah Evaluasi Formatif Dan Sumatif Dalam Kurikulum Merdeka. *Dewantara : Jurnal Pendidikan Sosial Humaniora*, *2*(4), 172–180. https://doi.org/10.30640/dewantara.v2i4.1783.

Gu, P. Y. (2021). An Argument-Based Framework for Validating Formative Assessment in the Classroom. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.605999

Gupta, K. (2023). Validity and Reliability of Students' Assessment: Case for Recognition as a Unified Concept of Valid Reliability. *International Journal of Applied & Basic Medical Research*, *13*(3), 129–132. https://doi.org/10.4103/ijabmr.ijabmr_382_23

Hartoko, G., & Leong, H. (2023). *Comparison Of Glove And Fasttext Algorithms On Cnn For Classification Of Indonesian News Categories*. *7*(1), 2023.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kernagaran, V., & Abdullah, A. (2025). *A Genre-Based Rubric For Assessing Higher-Order Thinking Skills In Student Essay Writing*. www.ijitsc.net

Khomsah, S., Ramadhani, R. D., & Wijayanto, S. (2022). The Accuracy Comparison Between Word2Vec and FastText On Sentiment Analysis of Hotel Reviews. *Jurnal RESTI*, *6*(3), 352–358. https://doi.org/10.29207/resti.v6i3.3711

Kloss, S., & Burdiles, G. (2024). Design and application of an instrument to evaluate argumentative academic essays. *Ogigia*, *36*, 257–288. https://doi.org/10.24197/ogigia.36.2024.257-288

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 757–770). International Committee on Computational Linguistics. https://aclanthology.org/2020.coling-main.66/

Kumar, V., & Boulanger, D. (2020). Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value. *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc.2020.572367

Kvasova, O., Kalinichenko, V., & Stus, V. '. (n.d.). Implications of training university teachers in developing local writing rating scales. In *Studies in Language Assessment* (Vol. 11). http://creativecommons.org/licenses/by/4.0/.

Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V., & Borodulin, A. (2024). *Applications of Long Short-Term Memory (LSTM) Networks in Polymeric Sciences: A Review*. https://doi.org/10.3390/polym

Maulani, G., Evenddy, S. S., Septiani, S., & Saptadi, N. T. S. (2024). *Ebook_EvaluasiPembelajaran_organized*. 1–329.

Milian, J. L. (2024). *Exploring Peer Assessment In Undergraduate Essay Writing: Chemistry Module Study* (Vol. 22). www.scientific-publications.net

Misgna, H., On, B. W., Lee, I., & Choi, G. S. (2025). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, *58*(2). https://doi.org/10.1007/s10462-024-11017-5

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2). https://doi.org/10.1016/j.rmal.2023.100050

Munaroh, N. L. (2024). Asesmen dalam Pendidikan : Memahami Konsep,Fungsi dan Penerapannya. *Dewantara : Jurnal Pendidikan Sosial Humaniora*, *3*(3), 281–297. https://doi.org/10.30640/dewantara.v3i3.2915

Pradani, K. A., & Suadaa, L. H. (2023). Automated Essay Scoring Menggunakan Semantic Textual Similarity Berbasis Transformer Untuk Penilaian Ujian Esai. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *10*(6), 1177–1184. https://doi.org/10.25126/jtiik.2023107338

Rajagede, R. A. (2021). Improving Automatic Essay Scoring for Indonesian Language using Simpler Model and Richer Feature. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 11–18. https://doi.org/10.22219/kinetik.v6i1.1196

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Rao, N. J., & Banerjee, S. (2023). Classroom Assessment in Higher Education. *Higher Education for the Future*, *10*(1), 11–30. https://doi.org/10.1177/23476311221143231

Sani, D. A., & Sarwani, M. Z. (2022). Koreksi Jawaban Esai Berdasarkan Persamaan Makna Menggunakan Fasttext dan Algoritma Backpropagation. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, *11*(2), 92–111. https://doi.org/10.23887/janapati.v11i2.49192

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306. https://doi.org/10.1016/j.physd.2019.132306

Sutriawan, Rustad, S., Shidik, G. F., & Pujiono. (2025). Performance Evaluation of Text Embedding Models for Ambiguity Classification in Indonesian News Corpus: A Comparative Study of TF-IDF, Word2Vec, FastText BERT, and GPT.

Ingenierie Des Systemes d'Information, *30*(6), 1469–1482. https://doi.org/10.18280/isi.300606

Uludag, P., & McDonough, K. (2022). Validating a rubric for assessing integrated writing in an EAP context. *Assessing Writing*, *52*. https://doi.org/10.1016/j.asw.2022.100609

Viñas, L. F. (2022). Testing the Reliability of two Rubrics Used in Official English Certificates for the Assessment of Writing. *Revista Alicantina de Estudios Ingleses*, *36*, 85–109. https://doi.org/10.14198/RAEI.2022.36.05

Waqas, M., & Humphries, U. W. (2024). A critical review of RNN and LSTM variants in hydrological time series predictions. In *MethodsX* (Vol. 13). Elsevier B.V. https://doi.org/10.1016/j.mex.2024.102946