BAB I

PENDAHULUAN

Penilaian merupakan komponen dasar dalam dunia pendidikan untuk mengukur pemahaman siswa. Proses penilaian secara manual yang selama ini digunakan dihadapkan pada tantangan fundamental. Untuk menjawab tantangan ini, perkembangan teknologi kecerdasan buatan dan Pemrosesan Bahasa Alami (NLP) memberikan alternatif melalui sistem penilaian otomatis. Bab ini akan menguraikan secara mendalam landasan yang mendasari penelitian ini, mulai dari latar belakang masalah, perumusan masalah yang spesifik, hingga tujuan, manfaat, dan batasan penelitian yang ditetapkan untuk mengevaluasi model yang akan dibuat dalam penelitian ini.

1.1 Latar Belakang

Dalam dunia pendidikan, evaluasi terhadap jawaban siswa merupakan bagian penting dalam mengukur tingkat pemahaman mereka terhadap suatu materi. Penilaian jawaban siswa secara manual oleh guru memiliki beberapa tantangan utama, seperti subjektivitas dalam penilaian, inkonsistensi skor antar penguji, serta waktu yang lama dalam proses koreksi. Seorang guru yang mengajar lebih dari satu kelas dengan jumlah siswa yang besar sering mengalami kesulitan dalam memberikan penilaian yang cepat dan akurat terhadap setiap jawaban siswa. Selain itu, variasi struktur kalimat dan penggunaan sinonim dalam jawaban siswa sering kali menyebabkan kesulitan dalam menentukan apakah suatu jawaban sudah memenuhi kriteria yang benar atau tidak.

Seiring dengan perkembangan teknologi kecerdasan buatan (*Artificial Intelligence*/AI) dan pemrosesan bahasa alami (*Natural Language Processing*/NLP), berbagai metode otomatisasi penilaian jawaban siswa telah dikembangkan. Salah satu pendekatan yang banyak digunakan adalah *Automated Essay Scoring* (AES), yang memanfaatkan teknik *machine learning* dan *deep learning* untuk mengevaluasi jawaban siswa secara otomatis. Penelitian oleh Wang (2024) menunjukkan bahwa pendekatan AES berbasis *semantic*, *thematic*, dan *linguistic representations* dapat meningkatkan akurasi penilaian jawaban berbasis teks, sehingga sistem dapat meniru cara manusia dalam memahami isi sebuah jawaban [1].

Berbagai penelitian lain juga telah mengembangkan model AES berbasis transformer untuk meningkatkan kualitas penilaian otomatis. Wangkriangkri et al. (2020) melakukan studi perbandingan antara berbagai *pretrained language models* seperti BERT, GPT, dan RoBERTa untuk sistem penilaian jawaban otomatis. Hasil penelitian menunjukkan bahwa model berbasis transformer memiliki keunggulan dalam menangkap hubungan semantik antar kata dalam jawaban siswa, dibandingkan dengan metode konvensional berbasis pencocokan kata kunci [2]. Selain itu, penelitian oleh Devlin et al. (2019) mengonfirmasi bahwa model BERT mampu memahami konteks kata dalam suatu kalimat secara *bidirectional*, sehingga lebih baik dalam memahami teks dibandingkan model berbasis RNN atau CNN [3].

Meskipun model berbasis transformer telah menunjukkan performa yang baik dalam berbagai tugas NLP, masih terdapat *research gap* dalam implementasinya untuk sistem penilaian otomatis jawaban siswa. Beseiso & Alzahrani (2020) menemukan bahwa meskipun BERT mampu memahami hubungan semantik dalam teks, model ini masih sensitif terhadap perbedaan struktur kalimat dan sering kali memberikan hasil yang tidak konsisten saat jawaban siswa menggunakan sinonim atau bentuk kalimat yang berbeda [4]. Penelitian lain oleh Nasreen et al. (2024) dalam konteks klasifikasi teks menemukan bahwa *deep learning* berbasis BERT dan teknik *feature selection* dapat meningkatkan akurasi model dalam menangani teks tidak terstruktur, tetapi masih mengalami kendala dalam memahami variasi frasa dalam jawaban panjang [5].

Di sisi lain, beberapa penelitian telah mencoba pendekatan lain untuk mengatasi keterbatasan model transformer dalam sistem penilaian otomatis. Kusumaningrum et al. (2024) mengembangkan model berbasis CNN-LSTM untuk menilai jawaban siswa dan menunjukkan bahwa kombinasi *Convolutional Neural Network* (CNN) dan *Long Short-Term Memory* (LSTM) mampu menangkap pola teks dengan lebih baik dibandingkan metode berbasis aturan. Namun, penelitian ini juga mengungkapkan bahwa model ini masih kesulitan dalam menangani variasi sinonim dalam jawaban siswa, yang menyebabkan hasil penilaian kurang akurat [6]. Sementara itu, penelitian oleh Kim et al. (2022) mengembangkan *chatbot* berbasis BERT untuk pencarian informasi, yang menunjukkan bahwa model transformer mampu menangkap konteks pertanyaan dengan lebih baik, tetapi masih mengalami kendala dalam memahami variasi struktur kalimat yang luas [7].

Untuk mengatasi *research gap* tersebut, penelitian ini mengusulkan pendekatan yang lebih fleksibel dengan menggabungkan GloVe-LSTM dengan metode ROUGE *Score*, TF-

IDF, dan *Cosine Similarity*. Model GloVe (*Global Vectors for Word Representation*) mampu merepresentasikan kata-kata dalam bentuk vektor dengan mempertimbangkan hubungan semantik antar kata, sementara LSTM (*Long Short-Term Memory*) digunakan untuk menangkap konteks kalimat secara lebih baik. Dengan kombinasi ini, model dapat memahami makna jawaban siswa meskipun terdapat perbedaan struktur kalimat atau penggunaan sinonim, sehingga lebih adaptif dalam sistem penilaian otomatis.

Selain itu, penelitian ini tidak hanya bergantung pada pencocokan kata kunci, tetapi juga mempertimbangkan makna keseluruhan jawaban siswa.

- *Cosine Similarity* digunakan untuk membandingkan vektor kalimat, bukan hanya kata individual, sehingga model dapat mengenali jawaban siswa yang memiliki makna serupa tetapi menggunakan kata berbeda.
- TF-IDF diterapkan bersama dengan *word embedding*, sehingga tidak hanya bergantung pada frekuensi kata, tetapi juga mempertimbangkan kemiripan antar kata menggunakan representasi vektor.
- ROUGE *Score* diterapkan untuk mengukur struktur kalimat dan kesamaan pola penulisan, meskipun siswa tidak menggunakan kata kunci secara eksplisit.

Studi oleh Onan & Alhumyani (2024) menemukan bahwa kombinasi *fuzzy topic modeling* dan jaringan transformer dapat meningkatkan efektivitas dalam ekstraksi teks akademik, yang menjadi referensi dalam penggunaan kombinasi metode dalam penelitian ini [8].

Dengan menggabungkan pendekatan GloVe-LSTM dan teknik evaluasi berbasis kemiripan teks, penelitian ini bertujuan untuk mengembangkan sistem penilaian otomatis yang lebih objektif, akurat, dan adaptif terhadap variasi jawaban siswa. Model yang dikembangkan diharapkan dapat mengatasi kekurangan yang ditemukan dalam penelitian sebelumnya, terutama dalam aspek konsistensi penilaian terhadap jawaban dengan struktur kalimat yang berbeda. Selain meningkatkan akurasi dalam sistem penilaian otomatis, penelitian ini juga memiliki potensi untuk digunakan dalam berbagai platform pendidikan digital, seperti *e-learning*, *Learning Management System* (LMS), atau sistem ujian daring. Dengan adanya sistem yang lebih cerdas dalam menilai jawaban siswa, diharapkan guru dapat mengurangi beban kerja dalam melakukan koreksi manual, sehingga dapat lebih fokus pada aspek pembelajaran yang lebih mendalam.

Dengan demikian, penelitian ini berkontribusi dalam pengembangan teknologi penilaian berbasis AI yang lebih akurat, efisien, dan adil bagi siswa. Diharapkan hasil dari penelitian ini dapat memberikan solusi inovatif dalam dunia pendidikan, khususnya dalam sistem evaluasi akademik berbasis *Natural Language Processing* (NLP) dan *machine learning*.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana kemampuan model penilaian otomatis berbasis GloVe-LSTM dalam menganalisis makna semantik jawaban siswa, khususnya pada kasus di mana jawaban tidak mengandung kata kunci secara eksplisit, dan faktor-faktor apa saja yang memengaruhi akurasi prediksi skornya?
2. Seberapa efektif model GloVe-LSTM dalam menangani variasi linguistik seperti struktur kalimat, penggunaan sinonim, dan gaya penulisan siswa, dan apa saja tantangan yang dihadapinya dalam konteks prediksi skor otomatis?
3. Bagaimana kontribusi dan pengaruh integrasi metode *Cosine Similarity*, TF-IDF, dan ROUGE *Score* dengan arsitektur GloVe-LSTM terhadap hasil prediksi skor untuk jawaban siswa dengan berbagai pola?
4. Bagaimana kinerja model prediksi skor yang dikembangkan ketika dievaluasi menggunakan metrik standar regresi (seperti MAE, RMSE, R^2 *Score*), metrik korelasi (*Pearson, Spearman*), serta metrik kesepakatan (seperti QWK jika skor dikategorikan), dan bagaimana responsnya terhadap jawaban yang mengandung sinonim serta variasi struktur kalimat?

1.3 Fokus Penelitian

Untuk menjawab rumusan masalah dan mencapai tujuan penelitian, maka penelitian ini akan difokuskan pada beberapa aspek berikut:

1. Pengembangan dan Analisis Model GloVe-LSTM
   - Mengimplementasikan *word embedding* GloVe untuk representasi kata dalam Bahasa Indonesia.
   - Membangun arsitektur *Long Short-Term Memory* (LSTM) untuk menangani konteks teks dan hubungan antar kata dalam jawaban siswa.

- Menganalisis kemampuan model dalam menangkap makna semantik dan menangani variasi linguistik.

2. Integrasi Algoritma Penilaian Teks
    - Menerapkan *Cosine Similarity* untuk mengukur kesamaan semantik berbasis vektor.
    - Menggunakan TF-IDF untuk mengidentifikasi dan membobot kata-kata kunci atau penting.
    - Menggunakan ROUGE *Score* untuk menilai kesamaan berdasarkan n-gram atau frasa.
    - Mengkaji pengaruh dan interaksi antara komponen LSTM dan algoritma-algoritma ini.
3. Evaluasi Kinerja Model
    - Membandingkan hasil prediksi skor model dengan skor manual (sebagai *ground truth*).
    - Mengevaluasi performa model menggunakan metrik regresi (MAE, RMSE, R^2 *Score*).
    - Mengukur korelasi (*Pearson* , *Spearman*) antara skor prediksi dan skor manual.
    - Menganalisis metrik kesepakatan (QWK) jika skor dikonversi menjadi kategori.
    - Menganalisis respons model pada skenario jawaban spesifik (misalnya, mengandung sinonim, struktur bervariasi, tidak relevan).
4. Validasi dan Uji Coba
    - Menggunakan dataset jawaban siswa berbahasa Indonesia yang relevan.
    - Menerapkan metode validasi yang sesuai (seperti *train-test split* atau *cross-validation*) untuk memastikan hasil evaluasi yang akurat dan dapat diandalkan.

1.4 Tujuan Penelitian

1 Untuk menganalisis kemampuan model penilaian otomatis berbasis GloVe-LSTM dalam menangkap makna semantik dari jawaban siswa, khususnya pada kasus di mana jawaban tidak mengandung kata kunci secara eksplisit, dan mengidentifikasi faktor-faktor yang memengaruhi akurasi prediksi skornya.

2 Untuk menyelidiki efektivitas dan tantangan yang dihadapi model GloVe-LSTM dalam menangani variasi struktur kalimat, penggunaan sinonim, dan perbedaan

gaya penulisan dalam jawaban siswa, khususnya dalam konteks prediksi skor otomatis.

3 Untuk mengkaji bagaimana integrasi metode *Cosine Similarity*, TF-IDF, dan ROUGE *Score* dengan arsitektur GloVe-LSTM berkontribusi atau memengaruhi hasil prediksi skor untuk jawaban siswa dengan berbagai pola.

4 Untuk mengevaluasi performa model prediksi skor yang dikembangkan berdasarkan metrik standar regresi (seperti MAE, RMSE, R^2 *Score*), metrik korelasi (*Pearson, Spearman*), serta metrik kesepakatan (seperti QWK jika skor dikategorikan), dan menganalisis respons model terhadap jawaban yang mengandung sinonim serta variasi struktur kalimat.

1.5 Manfaat Penelitian

1. Manfaat Teoritis
   - Memberikan pemahaman yang lebih mendalam mengenai penerapan dan tantangan model GloVe-LSTM yang diintegrasikan dengan algoritma ROUGE *Score*, TF-IDF, dan *Cosine Similarity* dalam memprediksi skor jawaban siswa berbahasa Indonesia, khususnya dalam aspek penangkapan makna semantik.
   - Menyajikan analisis mengenai kontribusi potensial dan interaksi antara komponen *deep learning* (GloVe-LSTM) dan metode evaluasi teks algoritmik dalam menghasilkan prediksi skor, yang dapat menjadi dasar untuk pengembangan arsitektur hibrida yang lebih efektif di masa depan.
   - Mengidentifikasi kekuatan dan keterbatasan pendekatan yang diteliti dalam menangani variasi linguistik (sinonim, struktur kalimat, kesalahan ejaan) dan jawaban tanpa kata kunci eksplisit, sehingga memberikan wawasan untuk penelitian NLP selanjutnya dalam konteks penilaian otomatis.
   - Menyumbangkan referensi metodologis dan hasil observasi awal dari pengembangan model penilaian otomatis spesifik untuk Bahasa Indonesia, yang dapat digunakan sebagai landasan atau pembanding bagi penelitian sejenis.
2. Manfaat Praktis
   - Menyediakan bukti konsep (*proof-of-concept*) pengembangan model yang mampu menghasilkan prediksi skor otomatis untuk jawaban siswa, yang berpotensi menjadi dasar untuk pengembangan alat bantu penilaian bagi

pendidik di masa depan, jika akurasi dan robustisitasnya dapat ditingkatkan lebih lanjut.

- Hasil analisis kinerja model pada berbagai skenario pengujian dapat memberikan masukan bagi perancang sistem penilaian otomatis mengenai aspek-aspek jawaban siswa yang masih sulit ditangani oleh teknologi saat ini dan memerlukan perhatian khusus.
- Meskipun belum dapat sepenuhnya menggantikan peran guru, skor prediksi yang dihasilkan model (jika divalidasi dan diinterpretasikan dengan hati-hati) berpotensi memberikan indikasi awal atau panduan sekunder kepada guru dalam proses evaluasi jawaban siswa dalam jumlah besar, sehingga dapat membantu mengarahkan fokus koreksi.
- Penelitian ini dapat mendorong eksplorasi lebih lanjut dalam pengembangan fitur umpan balik otomatis yang lebih detail, di mana prediksi skor dari model dapat menjadi salah satu komponen *input* untuk sistem umpan balik tersebut.

1.6 Batasan Penelitian

Agar penelitian ini tetap fokus dan menghasilkan hasil yang sesuai dengan tujuan yang telah ditetapkan, beberapa batasan berikut diterapkan:

1. Jenis Jawaban yang Dinilai
    - Penelitian ini hanya berfokus pada jawaban teks tertulis dalam bentuk esai atau uraian singkat.
    - Jawaban dalam bentuk gambar, tabel, rumus matematika, atau simbol non-teks tidak termasuk dalam cakupan penelitian ini.
2. Metode yang Digunakan
    - Model yang dikembangkan menggunakan GloVe-LSTM untuk memahami hubungan semantik dalam teks.
    - Pendekatan *Cosine Similarity*, TF-IDF, dan ROUGE *Score* digunakan untuk menilai kesesuaian jawaban siswa dengan kata kunci yang diberikan guru.
    - Model Transformer seperti BERT atau GPT tidak digunakan karena penelitian ini berfokus pada pendekatan berbasis *word embedding* dan *sequence modeling*.
3. Dataset yang Digunakan

- Dataset terdiri dari jawaban siswa yang telah dikumpulkan dari ujian atau tugas akademik, dengan kategori benar, kurang tepat, dan salah.
- Penelitian ini tidak menggunakan dataset dalam bahasa selain Bahasa Indonesia.
- Labeling data dilakukan secara manual oleh guru atau tenaga ahli untuk memastikan validitas data pelatihan.

4. Evaluasi dan Validasi Model
   - Model diuji menggunakan metrik evaluasi yang lebih tepat seperti precision, recall, F1-*Score*, ROUGE *Score*, *Cosine Similarity*, dan QWK. Metrik ini memungkinkan model untuk menilai jawaban berdasarkan kesamaan semantik dan struktur kalimat.
   - Perbandingan performa dilakukan dengan model *baseline*, seperti metode berbasis kata kunci sederhana (TF-IDF) atau *rule-based scoring*.
   - Uji coba model dilakukan dalam lingkungan pengujian berbasis *offline,* tanpa integrasi langsung dengan sistem *e-learning* atau *Learning Management System* (LMS).
5. Impementasi dan Penggunaan Model
   - Model ini dikembangkan dalam lingkungan pemrograman berbasis Python dengan pustaka *TensorFlow* atau PyTorch.
   - Model ini hanya diuji dalam simulasi berbasis dataset, sehingga belum diterapkan dalam sistem penilaian real-time yang digunakan secara luas dalam pendidikan formal

Dengan adanya batasan ini, penelitian tetap terarah pada pengembangan dan evaluasi model NLP berbasis GloVe-LSTM untuk penilaian otomatis jawaban siswa, tanpa memperluas cakupan ke implementasi sistem berbasis web atau aplikasi pendidikan secara langsung.