JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9317
Hal 967-973
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Implementasi Model Gpt-3.5 Turbo Untuk Otomatisasi Penilaian Esai Pada Sistem Pembelajaran Daring

Ade Suryadi[1], Sandra Jamu Kuryanti[1*], Cep Adiwihardja[2], Khaila Anjani[3], Meutya Febi Santoso [4]

[1]Fakultas Teknik dan Informatika, Prodi Sistem Informasi, Universitas Bina Sarana Informatika, DKI Jakarta, Indonesia
[2]Fakultas Teknik dan Informatika, Prodi Rekayasa Perangkat Lunak, Universitas Bina Sarana Informatika, DKI Jakarta, Indonesia
[3]Fakultas Ekonomi dan Bisnis, Prodi Manajemen, Universitas Bina Sarana Informatika, DKI Jakarta, Indonesia
[4]Fakultas Teknik dan Informatika, Prodi Sistem Informasi, Universitas Bina Sarana Informatika, DKI Jakarta, Indonesia
Email: [1]ade.axd@bsi.ac.id, [2*]sandra.sjk@bsi.ac.id, [3]cep.caw@bsi.ac.id, [4]khailaanjani5@gmail.com, [5]meutyafebis27@gmail.com
Email Penulis Korespondensi: sandra.sjk@bsi.ac.id
Submitted **04-08-2025;** Accepted **11-12-2025;** Published **31-12-2025**

Abstrak

Penilaian esai dalam pembelajaran daring membutuhkan waktu, tenaga, dan konsistensi yang sering kali sulit dijaga apabila dilakukan secara manual. Penelitian ini mengeksplorasi penggunaan model bahasa besar GPT-3.5 Turbo sebagai inti dari sistem otomatisasi penilaian esai untuk platform pembelajaran daring. Penelitian ini menggunakan pendekatan *Research and Development* (R&D) dengan model pengembangan ADDIE yang mencakup tahap *Analysis, Design, Development, Implementation*, dan Evaluation. Untuk kerangka penelitiannya sendiri mengadopsi dari kerangka kerja *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Sistem penilaian esai otomatis menggunakan *Prompt* 4 menunjukkan tingkat akurasi dan keandalan yang sangat tinggi. Model ini berhasil mencapai akurasi sebesar 94,3%, *F1-Score* 0,955, dan nilai *Cohen's Kappa* 0,878. Nilai Kappa yang tinggi ini menandakan adanya tingkat kesepakatan yang sangat kuat antara penilaian AI dengan *gold standard* yang divalidasi oleh guru, yang jauh melampaui kesepakatan awal antar guru itu sendiri yang hanya 0,1157. Kinerja superior Prompt 4 juga dikonfirmasi melalui nilai *Mean Absolute Error* (MAE) terendah sebesar 30,54 dan *Area Under the Curve* (AUC) tertinggi sebesar 0,956

Kata Kunci: GPT-3.5 Turbo; Penilaian Otomatis; Esai; Pembelajaran Daring

Abstract

Essay assessment in online learning requires significant time, effort, and consistency, which can be challenging to maintain when conducted manually. This study explores the use of the large language model GPT-3.5 Turbo as the core of an automated essay scoring system for online learning platforms. Employing a Research and Development (R&D) approach with the ADDIE development model—comprising Analysis, Design, Development, Implementation, and Evaluation phases—the research adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework for its methodology. The automated essay scoring system utilizing Prompt 4 demonstrated exceptionally high accuracy and reliability. The model achieved an accuracy of 94.3%, an F1-Score of 0.955, and a Cohen's Kappa value of 0.878. This high Kappa value indicates a very strong agreement between AI-generated assessments and the gold standard validated by educators, surpassing the initial inter-rater agreement among educators themselves, which was only 0.1157. The superior performance of Prompt 4 is also confirmed by the lowest Mean Absolute Error (MAE) of 30.54 and the highest Area Under the Curve (AUC) of 0.956.

Keywords: GPT-3.5 Turbo; Automated Scoring; Essay; Online Learning

1. PENDAHULUAN

Pembelajaran daring adalah pembelajaran yang dilakukan secara virtual menggunakan aplikasi perangkat lunak kompuer yang tersedia. Pembelajaran daring bukan hanya sekedar materi pelajaran yang dipin dahkan ke media internet dan bukan juga sekedar tugas-tugas yang dikirimkan melalui aplikasi pen dukung pembelajaran. Namun, pembelajaran daring harus tetap memperhatikan kompetensi yang ingin dicapai. [1]

Ujian esai merupakan evaluasi pembelajaran dalam bentuk soal esai yang mempunyai jawaban lebih bervariasi dibandingkan soal pilihan ganda[2]. Variasi jawaban tersebut memberikan kesulitan tersendiri bagi guru dalam menilai jawaban. Sistem penilaian esai dibangun untuk menjadi salah satu solusi yang dapat mempercepat dan mempermudah proses penilaian. Sistem penilaian esai pada penelitian ini dilakukan dengan mengukur kesamaan jawaban siswa dan kunci jawaban guru.[3]

Ujian esai *online* merupakan ujian yang menggunakan metode online dan mewajibkan siswa menjawab dengan kalimat mereka sendiri.[4] Penilaian dalam proses pembelajaran sangatlah penting karena dalam pengolahan data nilai menjadi salah satu pilar yang penting. Pengolahan data nilai yang baik akan menghasilkan nilai raport sebagai hasil evaluasi yang baik pula. [5]

Tujuan pembelajaran adalah langkah awal yang harus dilaksanakan perencanaan program dalam kegiatan pembelajaran, sehingga tujuan pembelajaran mencerminkan cita-cita setiap individu atau masyarakat yang dimiliki masing-masing. [6]

Ujian atau ulangan merupakan bagian penting dari proses evaluasi pembelajaran yang bertujuan untuk mengukur capaian kompetensi siswa. Evaluasi pembelajaran adalah tahapan yang sangat penting karena dapat menunjukkan sejauh mana keberhasilan proses pembelajaran telah tercapai. Bentuk evaluasi tidak hanya terbatas pada soal pilihan ganda (tes objektif), tetapi juga mencakup tes esai atau uraian. Tes esai merupakan bentuk pertanyaan yang mengharuskan siswa menjawab secara naratif, menjelaskan, dan memberikan argumen terhadap suatu topik Jawaban esai mencerminkan proses berpikir kritis siswa serta kualitas argumentasi dan logika yang digunakan. Berpikir kritis adalah kemampuan

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9317
Hal 967-973
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

berpikir secara sistematis dan terarah, membutuhkan pengetahuan dasar terhadap strategi pemecahan masalah, dan berperan penting dalam menumbuhkan kepercayaan diri siswa untuk mengembangkan ide, menganalisis, dan mengkomunikasikan argumen secara tertulis, keterampilan ini tidak muncul secara instan, melainkan membutuhkan latihan dan penguasaan teknik menulis yang baik, karena berpikir kritis dan menulis esai melibatkan pemikiran tingkat tinggi dan rasional, keduanya saling terkait erat.

Automated essay scoring (AES) merupakan salah satu penyelesaian yang ditawarkan untuk masalah tersebut. Definisi dari AES adalah sebuah tugas dari machine learning dalam NLP, dimana kita menciptakan suatu model yang bisa digunakan untuk memberikan nilai pada jawaban siswa yang berbentuk esai secara otomatis [7]

Automated Essay Scoring (AES) merupakan salah satu fitur *E-learning* yang dapat memudahkan guru dalam mencocokkan jawaban esai siswa dengan kunci jawaban. *Text similarity* adalah metode searching pencocokan kata dengan matching text berdasarkan kondisi (term) yang telah ditentukan. [8] Bahasa yang digunakan pada AES ini adalah bahasa Indonesia yang mempunyai morfologi kata yang berbeda dengan bahasa lainnya. Hasil yang didapatkan pada penelitian ini yaitu penggunaan *term frequency* (tf) lebih rendah dibandingkan dengan menggunakan inverse dokumen frequency (idf) dengan akurasi yang tinggi. [9] Manfaat penggunaan sistem penilaian jawaban esai otomatis yaitu sistem dapat memeriksa jawaban esai lebih detail dibandingkan manusia, sistem lebih konsisten dalam melakukan proses penilaian, dan penilaian sistem bersifat objektif tidak terpengaruh apapun. [10]

Oleh karena itu, esai menjadi metode evaluasi yang efektif untuk mengukur kemampuan berpikir kritis dan penguasaan konsep siswa secara lebih komprehensif. Namun demikian, penilaian esai menghadirkan tantangan tersendiri, terutama pada pembelajaran daring di mana guru dihadapkan pada keterbatasan sumber daya manusia, waktu, dan perangkat teknologi. Penilaian manual atas jawaban esai membutuhkan proses membaca dan menganalisis satu per satu secara mendalam, serta sering kali terpengaruh oleh kondisi fisik dan mental penilai yang kelelahan. Subjektivitas, inkonsistensi, dan tulisan siswa yang sulit dibaca menjadi hambatan utama dalam proses ini.

Masalah lain yang muncul adalah rendahnya literasi digital sebagian guru, terutama dari kalangan pendidik senior, yang belum terbiasa menggunakan perangkat lunak atau sistem digital dalam melakukan koreksi esai secara daring. Koreksi manual yang mengandalkan tenaga manusia tidak hanya memakan waktu, tetapi juga dapat menurunkan keakuratan penilaian. Situasi ini mendorong perlunya sistem penilaian esai yang bersifat otomatis, objektif, efisien, dan mudah dioperasikan untuk mendukung guru dalam melakukan evaluasi berbasis teks.

Kemajuan teknologi kecerdasan buatan (*Artificial Intelligence*/AI), khususnya dalam bidang pemrosesan bahasa alami (*Natural Language Processing*/NLP), membuka peluang besar untuk mengembangkan sistem penilaian otomatis. Model bahasa besar (*Large Language Models*/LLM) seperti GPT-3.5 Turbo telah menunjukkan kemampuan memahami konteks, makna, dan struktur argumentasi dalam bahasa alami dengan sangat baik. Penelitian Mahande et al. menunjukkan bahwa penggunaan model IndoBERT dan IndoSBERT dalam sistem *Automated Essay Scoring (AES)* dapat meningkatkan akurasi penilaian esai berbahasa Indonesia melalui pendekatan *transfer learning* dan embedding kalimat. Kemampuan LLM dalam memahami konteks semantik membuatnya sangat potensial digunakan dalam pengembangan sistem penilaian otomatis esai siswa.

Beberapa penelitian terkait penelitian yang dibuat, antara lain : Penelitian Eka, dkk tentang analisis metode cosine similarity aplikasi ujian online esai otomatis (studi kasus JTO Polinema) dijelaskan bahwa metode *Cosine Similarity* dijadikan sebagai pedoman dalam penelitian karena memiliki hasil kemiripan kata yang tepat. Ditinjau dari hasil nilai rata-rata *precisioncosine similarity* 93%, *recall cosine similarity* 86%, dan *f-measure cosine similarity* 89%, Dimana pengujian akurasi metode dilakukan pengujian *precision, recall*, dan *f-measure* dan berdasarkan hasil analisis dengan menggunakan metode yang telah dicoba diperoleh rata-rata 81%. [4]

Penelitian kharisma (2023) tentang automated essay scoring menggunakan semantic textual similarity berbasis tranformer untuk penilaian ujian esai, dimana berdasarkan penelitiannya disimpulkan bahwa dataset telah berhasil dibangun dengan melakukan proses perekaman esai, kunci jawaban, dan nilai pada lembar jawaban mahasiswa. Kemudian telah dibangun model berbasis *Transformers* menggunakan *fine-tuned IndoBERT* untuk menghitung semantic textual similarity dan model pembanding dengan TF-IDF yang dikombinasikan dengan cosine similarity dan linear regression. Dari ketiga model tersebut, diperoleh bahwa model *fine tuned IndoBERT* merupakan model terbaik yang dapat diaplikasikan untuk menilai esai secara otomatis pada ujian esai di Politeknik Statistika STIS. [7]

Penelitian berikut aristejo tentang sistem penggunaan chatgot dalam otomatisasi penilaian jawaban esai mahasiswa jurusan Teknik informatika di STMIK Antar Bangsa, dimana kecerdasan buatan dapat digunakan untuk mendukung pendidikan khususnya untuk penilaian esai, namun tetap diperlukan pendekatan yang bersifat kolaboratif antara teknologi dan tenaga pengajar agar kualitas evaluasi tetap terjaga secara optimal. Penilaian dari soal dan jawaban yang sama namun dilakukan pada waktu yang berbeda, maka ChatGPT dapat memberikan perbedaan hasil penilaian yang signifikan, bahkan hingga ke tingkat komentar dan saran, dosen sebagai pemegang peran penting sebagai validator dan penentu Keputusan akhir dalam proses penilaian esai mahasiswa. [11]

Penelitian oleh Theo, dkk tentang implementasi kecerdasan buatan chatgpt dalam pembelajaran tentang ChatGPT merupakan sistem kecerdasan buatan AI yang memungkinkan interaksi melalui percakapan berbasis teks. Fungsionalitas ChatGPT dalam konteks pembelajaran melibatkan penerjemahan bahasa, pemberian rekomendasi, peningkatan produktivitas, peran sebagai sumber belajar interaktif, serta bantuan dalam menyelesaikan tugas dan memecahkan masalah bagi siswa, namun ChatGPT memiliki beberapa kelemahan dalam pembelajaran, seperti keterbatasan

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9317
Hal 967-973
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

kontekstual, kurangnya koneksi emosional, ketidakmampuan membedakan fakta dan opini, jawaban yang tidak selalu tepat, dan kurangnya interaksi dengan manusia.[12]

Penelitian ini bertujuan mengembangkan sistem penilaian otomatis berbasis aplikasi mobile Android dengan memanfaatkan model GPT-3.5 Turbo yang diakses melalui *Application Programming Interface* (API). Sistem ini akan dirancang untuk memberikan penilaian biner ("Benar" atau "Salah") dengan konfigurasi teknis seperti *temperature* 0.0 untuk menjaga konsistensi respons dan batasan *token* agar hasil tetap akurat dan relevan. Metode yang digunakan dalam penelitian ini adalah *Research and Development* (R&D) dengan pendekatan ADDIE, yang meliputi tahapan analisis kebutuhan, desain sistem, pengembangan, implementasi, dan evaluasi performa sistem. Penilaian otomatis yang dihasilkan akan dibandingkan dengan hasil penilaian manual oleh guru menggunakan pendekatan kuantitatif melalui uji korelasi antar skor, serta pendekatan kualitatif terhadap kualitas output sistem.

Kebaruan dari penelitian ini terletak pada pemanfaatan GPT-3.5 Turbo sebagai mesin penilai esai berbasis mobile dalam konteks pendidikan Indonesia. Berbeda dengan penelitian sebelumnya yang banyak berfokus pada model lokal seperti IndoBERT, penelitian ini menguji efektivitas LLM generatif global melalui integrasi API langsung ke dalam aplikasi Android. Selain itu, fokus pada skenario penilaian biner juga merupakan pendekatan baru untuk mengurangi ambiguitas penilaian dan meningkatkan kepraktisan penggunaannya di sekolah menengah atas.

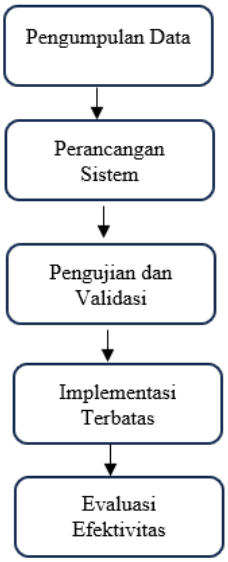
State of the art dari penelitian ini berada pada irisan antara teknologi NLP, pendidikan digital, dan pengembangan sistem penilaian berbasis AI. Meski sudah ada studi tentang AES di berbagai negara, khususnya dalam bahasa Inggris, penerapan GPT-3.5 Turbo untuk konteks bahasa Indonesia dan evaluasi esai dalam skala pendidikan menengah masih sangat terbatas

Efektivitas GPT-3.5 dalam situasi yang memerlukan penilaian otomatis dapat meningkat secara signifikan dengan penyesuaian kecil. GPT-3.5 dapat dimodifikasi untuk menangkap nuansa dan kompleksitas dari kriteria penilaian dengan mengubah bobot model berdasarkan dataset penilaian tertentu, sehingga menghasilkan keselarasan yang lebih baik dengan ekspektasi penilai manusia. GPT-3.5 merupakan kandidat yang kuat untuk pengembangan sistem penilaian otonom karena arsitektur dan kemampuannya. GPT-3.5-Turbo merupakan sistem penilaian otomatis yang lebih akurat dan konsisten, serta merupakan model yang memahami aspek-aspek kompleks dari subjek atau tugas tertentu. [9]

ChatGPT 3.5 Turbo dan GPT-4, telah menunjukkan kemampuan luar biasa dalam memahami, menghasilkan, dan memanipulasi bahasa manusia, [13]

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan *Research and Development* (R&D) dengan model pengembangan ADDIE yang mencakup tahap *Analysis, Design, Development, Implementation*, dan Evaluation. ADDIE merupakan desain instruksional yang berfokus pada individu, memiliki fase yang langsung dan berjangka panjang, bersifat sistematis, dan menggunakan pendekatan sistem dalam memahami pengetahuan dan pembelajaran manusia [14]. Salah satu bentuk penerapan model ADDIE dalam pengembangan produk adalah pada pengembangan bahan ajar. Dimana model ini dapat berperan untuk mendukung penyusunan kerangka teoritis dalam perancangan pembelajaran. Model ADDIE dirancang sebagai proses berurutan, tetapi fleksibel yang dapat disesuaikan dengan kebutuhan proyek, karena model ini tersusun secara sistematis dan terprogram. Adapun tahapan penelitian yang digunakan pada penelitian ini adalah sebagai berikut :



Gambar 1. Tahapan Metode Penelitian

Dari gambar 1 diatas dapat dijelaskan bahwa :

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9317
Hal 967-973
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

a. Pengumpulan Data
   Data dikumpulkan dari hasil asesmen guru terhadap jawaban esai siswa. Sebanyak 50 esai diambil dari mata pelajaran Bahasa Indonesia dengan topik opini, deskripsi, atau analisis sosial. Setiap esai telah dinilai secara manual oleh guru. Data ini digunakan untuk membandingkan hasil koreksi manual dengan hasil penilaian otomatis berbasis API.
b. Perancangan Sistem
   Peneliti menyusun *prompt engineering* untuk mengatur model GPT-3.5 Turbo agar hanya memberikan respons berupa "Benar" atau "Salah" terhadap jawaban siswa. Prompt diformat sedemikian rupa dengan menyisipkan soal, jawaban siswa, dan kunci jawaban. Parameter temperature disetel ke 0 dan max_tokens dibatasi agar hasil tetap konsisten. Implementasi dilakukan melalui pemrograman Dart, PHP, dan cURL untuk memanggil API secara *real-time*.
c. Pengujian dan Validasi
   Hasil penilaian otomatis dibandingkan dengan hasil penilaian guru untuk mengukur tingkat kesesuaian. Penilaian dilakukan dalam bentuk klasifikasi biner. Validasi dilakukan secara kuantitatif menggunakan akurasi dan persentase kesesuaian antara hasil API dan guru. Selain itu, dilakukan diskusi dengan guru untuk mengevaluasi kemanfaatan dan kepraktisan sistem.
d. Implementasi Terbatas
   Sistem diujicobakan secara langsung kepada guru dan siswa selama dua minggu. Guru diminta mengunggah jawaban siswa melalui aplikasi Android dan menerima hasil koreksi dari sistem secara otomatis. Waktu yang dibutuhkan serta tingkat kepuasan guru terhadap sistem dievaluasi melalui angket dan wawancara sederhana.
e. Evaluasi Efektivitas
   Efektivitas sistem dinilai berdasarkan tiga indikator: (a) penghematan waktu koreksi oleh guru, (b) tingkat kesesuaian hasil koreksi dengan penilaian manual minimal 80%, dan (c) tingkat kepuasan guru terhadHap penggunaan sistem. Data dianalisis secara deskriptif dan kuantitatif sederhana.

Untuk kerangka penelitiannya sendiri mengadopsi dari kerangka kerja *Cross-Industry Standard Process for Data Mining* (CRISP-DM) untuk memandu proses pengembangan sistem penilaian esai otomatis. Metodologi ini dipilih karena sifatnya yang sistematis, fleksibel, dan iteratif, sehingga sangat sesuai untuk proyek berbasis data seperti ini [15]. CRISP-DM menyediakan enam tahapan yang jelas dan terstruktur yang memungkinkan penyesuaian terhadap berbagai jenis data dan tujuan penelitian, serta mendukung perbaikan model secara berkelanjutan berdasarkan hasil evaluasi.[16]
CRISP-DM merupakan sebuah metodologi yang menyediakan pendekatan umum untuk membentuk dan merencanakan proyek data mining.[17]. 6 tahapan CRISP-DM antara lain, yaitu *business understanding, data understanding, data preparation, modeling, evaluation, dan deployment* [18]

3. HASIL DAN PEMBAHASAN

3.1 Hasil *Business Understanding*

Hasil business understanding merupakan tahapan pertama dalam Data Mining untuk mendefinisikan permasalahan yang akan dibahas, menyiapkan strategi apa yang akan digunakan dan juga menentukan tujuan penelitian atau planning yang telah dilakukan untuk mencapai sesuatu yang dimaksud. [19]

Penilaian pada soal ujian esai memerlukan waktu yang lebih lama dibanding menilai soal yang berbentuk objektif atau pilihan ganda. Selain itu, penilaian ini juga berpotensi menimbulkan masalah konsistensi dan ketepatan penilaian jawaban, baik dengan penilai yang sama maupun berbeda.

Oleh sebab itu, penelitian ini berfokus untuk mengatasi masalah tersebut, dengan tujuan utama untuk mengembangkan serta mengevaluasi sistem penilaian esai otomatis berbasis GPT-3.5 Turbo, sehingga bisa mengurangi beban kerja guru dan bisa menyediakan masukan yang cepat dan akurat pada siswa. Untuk mencapai tujuan tersebut, adapun sasaran yang ditargetkan dalam penelitian ini yaitu merancang dan membandingkan berbagai strategi prompt engineering, mengukur tingkat akurasi setiap prompt terhadap penilaian guru ahli sebagai gold standard, serta mengidentifikasi prompt paling efektif beserta ambang batas skor optimalnya. Tingkat keberhasilan atau kriteria kesuksesan dalam penelitian ini dapat diukur dari hasil pencapaian F1-*Score* dan *Cohen's Kappa* yang tinggi, antara *prompt* terbaik dengan *gold label*, sehingga menunjukkan adanya tingkat kesepakatan yang kuat.

Permasalahan Sistem ujian online saat ini Adalah guru harus menilai secara manual untuk soal-soal kategori esai. Permasalahan ini tentu menyebabkan delay dalam memberikan hasil kepada siswa, sedangkan soal-soal kategori pilihan ganda sudah tentu dapat memberikan hasil cepat, namun dikarenakan soal tersebut ada yang pilihan ganda dan esai maka total nilai tidak langsung dimunculkan menunggu koreksi soal esai dari guru. Jika permasalahan koreksi jawaban esai ini dapat diatasi dengan penilai otomatis maka akan banyak memberikan manfaat seperti waktu tunggu siswa menjadi tidak ada, efisiensi biaya koreksi guru, efisiensi beban guru dan meningkatkan kualitas sekolah.

3.2 Hasil Data *Understanding*

Data *understanding* merupakan sebuah tahapan dalam metodologi ilmu data dan pengembangan kecerdasan buatan yang bertujuan untuk mendapatkan pemahaman awal tentang data yang diperlukan untuk memecahkan masalah bisnis tertentu.[20]

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9317
Hal 967-973
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Penelitian menggunakan data yang merupakan hasil ujian siswa. Data primer yang digunakan merupakan hasil ujian harian Bersama (UHB) mata pelajaran Bahasa Indonesia kelas XI TKJ 1 dan XI TKJ 2 SMK Muhammadiyah Bumiayu pada tanggal 25 Agustus – 5 September 2025, yang terdiri dari lima butir soal esai, serta kunci jawaban guru. Secara keseluruhan, data mentah terdiri dari 210 baris dengan 5 soal yang dikerjakan oleh 42 siswa.
Adapun Soal yang digunakan diantaranya:

a. Jelaskan perbedaan mendasar antara paragraf deduktif dan induktif, dan berikan satu contoh paragraf deduktif. Kunci Jawaban: Perbedaan mendasar antara paragraf deduktif dan induktif terletak pada letak kalimat utamanya. Paragraf deduktif menempatkan kalimat utama di awal paragraf, yang kemudian diikuti oleh kalimat-kalimat penjelas. Polanya bergerak dari hal yang umum ke hal yang khusus. Sebaliknya, paragraf induktif menempatkan kalimat utama di akhir paragraf, diawali oleh kalimat-kalimat penjelas atau data khusus, lalu ditutup dengan kesimpulan sebagai kalimat utama di bagian penutup. Polanya bergerak dari hal yang khusus ke hal yang umum.

 Contoh Paragraf Deduktif:
 Bunga mawar adalah tanaman hias yang sangat populer di seluruh dunia. Tanaman ini dikenal karena keindahan bunganya yang beragam warna, seperti merah, putih, kuning, dan oranye. Selain itu, mawar juga memiliki aroma harum dan sering dijadikan simbol cinta. Oleh karena itu, bunga mawar banyak dibudidayakan sebagai tanaman hias di pekarangan, rumah, maupun di perkebunan.
b. Mengapa penting untuk dapat membedakan antara paragraf deduktif dan induktif saat membaca sebuah teks? Kunci Jawaban: Memahami perbedaan kedua jenis paragraf ini sangat penting untuk meningkatkan kemampuan membaca secara efektif. Dengan mengenali polanya, pembaca dapat dengan cepat menemukan ide pokok atau gagasan utama sebuah tulisan. Ini memungkinkan pembaca untuk memahami inti dari paragraf tanpa harus membaca seluruh detailnya. Hal ini sangat berguna dalam kegiatan membaca cepat, merangkum, atau menganalisis teks, karena fokus utama langsung tertuju pada poin yang paling penting.
c. Jelaskan perbedaan mendasar antara pernyataan fakta dan opini. Berikan satu contoh untuk masing-masing jenis pernyataan tersebut!. Kunci Jawaban: Perbedaan mendasar antara fakta dan opini terletak pada sifat kebenarannya. Pernyataan fakta adalah pernyataan yang dapat dibuktikan kebenarannya, bersifat objektif, dan datanya bisa diverifikasi. Sementara itu, pernyataan opini adalah pandangan, pendapat, atau penilaian pribadi seseorang yang bersifat subjektif dan belum tentu benar. Contoh Fakta: “Indonesia adalah negara kepulauan terbesar di dunia.” (Dapat diverifikasi melalui data geografis). Contoh Opini: “Pelajaran Bahasa Indonesia adalah pelajaran yang paling sulit.” (Merupakan pendapat pribadi yang bisa berbeda bagi setiap orang)
d. Dalam sebuah teks laporan ilmiah atau berita, pernyataan fakta lebih diutamakan daripada pernyataan opini. Jelaskan mengapa hal ini penting dan apa dampak jika sebuah teks laporan ilmiah lebih didominasi oleh opini!. Kunci Jawaban: Pernyataan fakta lebih diutamakan karena laporan ilmiah atau berita bertujuan untuk menyampaikan informasi yang objektif, akurat, dan dapat dipercaya. Fakta memberikan landasan yang kuat dan meyakinkan bagi pembaca.
e. Buatlah satu pernyataan fakta dan satu pernyataan opini dengan topik “Manfaat media sosial”. Kunci Jawaban: Contoh Fakta: “Indonesia adalah negara kepulauan terbesar di dunia.” (Dapat diverifikasi melalui data geografis). Contoh Opini: “Pelajaran Bahasa Indonesia adalah pelajaran yang paling sulit.” (Merupakan pendapat pribadi yang bisa berbeda bagi setiap orang). Pernyataan Fakta: “Hingga tahun 2024, pengguna aktif media sosial di Indonesia mencapai lebih dari 190 juta orang.” (Data ini dapat diverifikasi dari berbagai sumber riset digital). Pernyataan Opini: “Media sosial merupakan sarana terbaik untuk meningkatkan kreativitas.” (Pernyataan ini adalah pandangan pribadi. Meskipun sebagian orang setuju, sebagian lain mungkin memiliki pendapat berbeda).

3.3 Hasil Data *Preparation*

Pada tahap ini dilakukan beberapa langkah untuk memastikan kualitas serta kesiapan data sebelum dianalisis lebih lanjut. Pertama, dilakukan validasi ahli dan reliabilitas antar penilai melalui penghitungan *Cohen’s Kappa* antara guru 1 dengan guru 2. Hasil perhitungan menunjukkan nilai Kappa sebesar 0.1157. Nilai ini mengindikasikan tingkat kesepakatan yang rendah (*Slight Agreement*). Temuan ini menjadi bukti kuantitatif yang menggarisbawahi masalah subjektivitas dalam penilaian manual, yang menjadi justifikasi utama penelitian ini. Langkah selanjutnya adalah pembuatan *Gold Standard Label* dengan total 210 data esai siswa, terdapat 127 data dengan kesepakatan atau adanya penilaian yang sama antara kedua guru dan adan 83 data yang menunjukkan perbedaan. Pada kolom “*gold_label*” digunakan sebagai hasil konsensus final yang menjadi acuan dalam evaluasi model. Dari total 210 data, terdapat 127 data (60.48%) yang disepakati dan 83 data yang berbeda penilaiannya. Perbedaan ini kemudian didiskusikan dan diselesaikan oleh kedua guru untuk mencapai konsensus, yang hasilnya menjadi kolom gold_label. Kolom inilah yang digunakan sebagai standar emas (*ground truth*) final.

Tahap berikutnya adalah pembersihan dan parsing data AI. Setiap respon teks yang dihasilkan dari masing-masing prompt diparsing menjadi kolom keputusan (benar/salah) dan skor (nilai numerik). Tahap akhir, yaitu dataset final dianalisis, yang terdiri dari 210 data valid. Dataset ini digunakan sebagai dasar untuk analisis selanjutnya, yaitu pemodelan dan evaluasi.

3.4 Hasil Pemodelan

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9317
Hal 967-973
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Teknik pemodelan yang digunakan pada penelitian ini adalah rekayasa prompt *(prompt engineering)* pada model GPT-3.5 Turbo. Proses pemodelan ini dirancang dengan sifat iteratif, dimana beberapa strategi prompt dirancang sebagai "model" yang berbeda. Dalam tahap ini, ada lima model atau prompt yang diuji.

3.5 Hasil Evaluasi

Dataset yang telah ditentukan kemudian dilakukan proses analisis kuantitatif. Setiap respons model dalam bentuk teks, misalnya *"Benar Skor: 80"*, diparsing menjadi dua komponen:

a. Label keputusan (*Benar/Salah*).
b. Skor numerik (0–100).

Kedua komponen ini kemudian digunakan dalam evaluasi. Label keputusan dianalisis dengan *confusion matrix* dan metrik klasifikasi (Akurasi, Presisi, Recall, F1, dan Kappa), sementara skor numerik digunakan untuk perhitungan *Mean Absolute Error* (MAE) dan *ROC-AUC*.

4. KESIMPULAN

Berdasarkan dari hasil analisis dan evaluasi yang telah dilakukan maka dapat ditarik Kesimpulan bahwa Berbagai strategi *prompt engineering* memberikan pengaruh yang signifikan terhadap kinerja model *GPT-3.5 Turbo* dalam menilai esai secara otomatis. Strategi yang paling efektif adalah *Prompt* 4, yang menggunakan pendekatan berbasis rubrik penilaian dengan pedoman skor dan batasan yang detail. *Prompt* ini terbukti secara statistik lebih unggul dibandingkan strategi lain, seperti yang ditunjukkan oleh hasil uji *McNemar d*engan *p-value* < *0.000001*. Sistem penilaian esai otomatis menggunakan *Prompt* 4 menunjukkan tingkat akurasi dan keandalan yang sangat tinggi. Model ini berhasil mencapai akurasi sebesar 94,3%, *F1-Score* 0,955, dan nilai *Cohen's Kappa* 0,878. Nilai Kappa yang tinggi ini menandakan adanya tingkat kesepakatan yang sangat kuat antara penilaian AI dengan *gold standard* yang divalidasi oleh guru, yang jauh melampaui kesepakatan awal antar guru itu sendiri yang hanya 0,1157. *Prompt* 4 diidentifikasi sebagai strategi terbaik dengan ambang batas skor optimal *(≥46)* yang konsisten dengan keputusan teks aslinya (*As-Is*). c, yang menunjukkan kemampuan diskriminatif terbaik dalam membedakan jawaban benar dan salah. Oleh karena itu, Prompt 4 merupakan model yang paling layak untuk diimplementasikan.

REFERENCES

[1] I. Sarto, "PENGARUH PEMBELAJARAN DARING TERHADAP MINAT BELAJAR SISWA PADA MASA PANDEMI COVID-19 KELAS V SDN CENDRAWASIH 1 MAKASSAR The Effect Of Online Learning On Students' Interest Learning During The Covid-19 Pandemic Of Class V SDN Cendrawasih 1 Makassar."
[2] N. L. Kinanti and A. Qoiriah, "Sistem Penilaian Otomatis Jawaban Esai Bahasa Indonesia Berdasarkan Kemiripan Kalimat Menggunakan Syntactic-Semantic Similarity," Journal of Informatics and Computer Science, vol. 02, 2020.
[3] J. Esai, M. Jurusan, T. Informatika, D. Stmik, and A. B. Aristejo, "Penggunaan ChatGPT dalam Otomatisasi Penilaian".
[4] E. L. Amalia1, A. J. Jumadi, I. A. Mashudi3, W. Wibowo4, and P. N. Malang, "ANALISIS METODE COSINE SIMILARITY PADA APLIKASI UJIAN ONLINE ESAI OTOMATIS (STUDI KASUS JTI POLINEMA) COSINE SIMILARITY METHOD ANALYSIS ON AUTOMATIC ESAI ONLINE TEST APPLICATION", doi: 10.25126/jtiik.202184356.
[5] A. Sumbaryadi and P. Christo, "SISTEM INFORMASI PENILAIAN HASIL BELAJAR SISWA SEKOLAH MENENGAH KEJURUAN (SMK) BERBASIS WEB," Sistem Informasi |, vol. 6, no. 1, pp. 48–53, 2019.
[6] J. Pendidikan dan Pengabdian kepada Masyarakat, N. Putri Mawarny, S. Holida, and N. Sari Siregar, "Tahun 2022 | Hal," vol. 1, no. 3, pp. 30–39, [Online]. Available: https://jurnal.permapendis-sumut.org/index.php/pema
[7] K. A. Pradani and L. H. Suadaa, "Automated Essay Scoring Menggunakan Semantic Textual Similarity Berbasis Transformer Untuk Penilaian Ujian Esai," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 10, no. 6, pp. 1177–1184, Dec. 2023, doi: 10.25126/jtiik.2023107338.
[8] R. Ahmad, "E-learning Automated Essay Scoring System Menggunakan Metode Searching Text Similarity Matching Text," Jurnal Penelitian Enjiniring, vol. 22, no. 1, pp. 38–43, May 2019, doi: 10.25042/jpe.052018.07.
[9] E. Latif and X. Zhai, "Fine-tuning ChatGPT for automatic scoring," Computers and Education: Artificial Intelligence, vol. 6, Jun. 2024, doi: 10.1016/j.caeai.2024.100210.
[10] N. L. Kinanti and A. Qoiriah, "Sistem Penilaian Otomatis Jawaban Esai Bahasa Indonesia Berdasarkan Kemiripan Kalimat Menggunakan Syntactic-Semantic Similarity," Journal of Informatics and Computer Science, vol. 02, 2020.
[11] J. Esai, M. Jurusan, T. Informatika, D. Stmik, and A. B. Aristejo, "Penggunaan ChatGPT dalam Otomatisasi Penilaian".
[12] "932+Template+JPT+2020+26862-26869".
[13] C. A. Mallio, C. Bernetti, A. C. Sertorio, and B. B. Zobel, "ChatGPT in radiology structured reporting: analysis of ChatGPT-3.5 Turbo and GPT-4 in reducing word count and recalling findings," Feb. 10, 2024, AME Publishing Company. doi: 10.21037/qims-23-1300.
[14] U. Sultan Syarif Kasim and K. Kunci, "Pengembangan Model ADDIE (Analisys, Design, Development, Implemetation, Evaluation)."
[15] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," Journal of Applied Informatics and Computing, vol. 5, no. 2, pp. 103–108, 2021, doi: 10.30871/jaic.v5i2.3200.
[16] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," 2021. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[17] A. Rianti et al., "CRISP-DM: Metodologi Proyek Data Science."
[18] I. Budiman et al., "Data Clustering Menggunakan Metodologi CRISP-DM Untuk Pengenalan Pola Proporsi Pelaksanaan Tridharma," 2011.
[19] F. T. Informasi, D. Komunikasi, D. Ratna, M. Nafisah, and A. Hendrawan, "SEMINAR NASIONAL INOVASI DAN TREN TEKNOLOGI (SINATTI) PENERAPAN METODE CRISP-DM DENGAN ALGORITMA K-MEANS CLUSTERING UNTUK ANALISA KEMISKINAN DAN KONSUMSI PER KAPITA DI JAWA TENGAH SELAMA PANDEMI."
[20] I. Gede Iwan Sudipa et al., DATA MINING. [Online]. Available: www.globaleksekutifteknologi.co.id