Vol. 7, No. 2, Agustus 2026, hal. 317-323

Jurnal Computer Science and Information Technology (CoSciTech)

p-ISSN: 2723-567X
e-ISSN: 2723-5661

http://ejurnal.umri.ac.id/index.php/coscitech/index

Model Deteksi Teks Generatif menggunakan Stacking Ensemble Berbasis DistilBERT dan Few-Shot Adaptation

Sriwahyuningsih Piu [*1], Usman [2], Arwansyah [3]
Email: [1]sri.wahyuningsih@undipa.ac.id, [2]usman@undipa.ac.id, [3]arwansyah@undipa.ac.id

[123]Teknik Informatika, Universitas Dipa Makassar

Diterima: 11 Juni 2026 | Direvisi: - | Disetujui: 25 Agustus 2026

Abstrak

Pesatnya perkembangan teknologi *Generative Artificial Intelligence* (AI) seperti ChatGPT, Claude, dan Gemini telah membawa tantangan baru bagi integritas akademik di lingkungan perguruan tinggi. Deteksi terhadap karya tulis yang dihasilkan oleh AI menjadi krusial untuk memastikan orisinalitas dan kejujuran intelektual mahasiswa. Namun, detektor AI konvensional sering kali bersifat statis dan memiliki tingkat akurasi yang rendah terhadap variasi gaya penulisan yang dinamis. Penelitian ini bertujuan untuk mengembangkan sistem deteksi teks generatif yang lebih adaptif dan akurat dengan menerapkan metode Stacking Ensemble Learning. Sistem ini mengintegrasikan empat model klasifikasi dasar (*Logistic Regression, SVM, MLP,* dan *Random Forest*) yang diperkuat dengan fitur semantik tingkat tinggi menggunakan DistilBERT Embeddings. Selain itu, sistem ini memanfaatkan Google Gemini API sebagai instrumen augmentasi fitur untuk mendapatkan penilaian risiko kualitatif secara *real-time*. Inovasi utama dalam penelitian ini terletak pada penerapan teknik Few-Shot Adaptation. Fitur ini memungkinkan dosen atau pengguna di perguruan tinggi untuk melakukan kalibrasi ulang pada *Meta-Classifier* secara instan hanya dengan mengunggah sedikit contoh dokumen asli (*human-written*) dan dokumen buatan AI (*AI-generated*). Sistem dikembangkan berbasis web menggunakan *framework* Gradio, yang mampu memproses berbagai format dokumen akademik seperti .pdf, .docx, dan .txt. Hasil penelitian ini diharapkan dapat memberikan kontribusi nyata bagi perguruan tinggi dalam bentuk instrumen digital yang mampu memvalidasi keaslian dokumen akademik secara presisi. Selain memberikan skor probabilitas, sistem juga menyediakan fitur Text Highlighting untuk menandai bagian teks yang mencurigakan, sehingga mempermudah proses evaluasi oleh tenaga pendidik.

Kata kunci: *Deteksi AI, Stacking Ensemble, DistilBERT, Few-Shot Adaptation, LLM.*

Generative Text Detection Model using DistilBERT-Based Stacking Ensemble and Few-Shot Adaptation

Abstract

The rapid development of Generative Artificial Intelligence (AI) technologies such as ChatGPT, Claude, and Gemini has brought new challenges to academic integrity in higher education. Detection of AI-generated written works has become crucial to ensure the originality and intellectual honesty of students. However, conventional AI detectors are often static and have low accuracy against dynamic variations in writing styles. This research aims to develop a more adaptive and accurate generative text detection system by applying the Stacking Ensemble Learning method. This system integrates four basic classification models (Logistic Regression, SVM, MLP, and Random Forest) enhanced with high-level semantic features using DistilBERT Embeddings. In addition, this system utilizes the Google Gemini API as a feature augmentation instrument to obtain real-time qualitative risk assessments. The main innovation in this research lies in the application of the Few-Shot Adaptation technique. This feature allows lecturers or users in higher education to instantly recalibrate the Meta-Classifier by simply uploading a few examples of original documents (human-written) and AI-generated documents. The system was developed web-based using the Gradio framework, which can process various academic document formats such as .pdf, .docx, and .txt. The results of this research are expected to provide a tangible contribution to higher education institutions in the form of a digital instrument

capable of precisely validating the authenticity of academic documents. In addition to providing a probability score, the system also includes a Text Highlighting feature to mark suspicious text sections, thus simplifying the evaluation process by educators.

Keywords: *AI Detection, Stacking Ensemble, DistilBERT, Few-Shot Adaptation, LLM.*

1. PENDAHULUAN

Transformasi digital dalam dunia pendidikan telah mencapai titik balik dengan hadirnya teknologi *Large Language Models* (LLM) seperti ChatGPT, Gemini, dan Claude. adopsi AI oleh mahasiswa dalam membantu proses penyusunan dokumen akademik—seperti laporan praktikum, esai, hingga skripsi—meningkat secara signifikan. Meskipun AI dapat menjadi alat bantu, penggunaan tanpa batasan yang jelas mengancam integritas akademik dan esensi penilaian kemampuan mahasiswa. Permasalahan utama terletak pada kemampuan LLM saat ini yang mampu menghasilkan teks dengan struktur koheren dan tata bahasa yang hampir sempurna, sehingga sulit dibedakan dengan tulisan manusia secara kasat mata. Perkembangan teknologi kecerdasan buatan telah melahirkan *Large Language Models* (LLM) seperti GPT-4, Claude, dan Gemini yang mampu menghasilkan teks dengan kualitas menyerupai tulisan manusia [1]. Model-model ini dilatih menggunakan dataset skala masif untuk memprediksi token selanjutnya dalam sebuah urutan, sehingga menghasilkan teks yang koheren secara sintaksis [2]. Namun, teks generatif sering kali memiliki karakteristik statistik yang berbeda, seperti rendahnya variasi kerumitan (*low perplexity*) dan pola keterpencaran yang seragam (*low burstiness*) dibandingkan tulisan manusia [3]. *Natural Language Processing* (NLP) merupakan bidang AI yang memfokuskan pada interaksi antara komputer dan bahasa manusia [4]. Salah satu terobosan penting adalah arsitektur Transformer yang memungkinkan representasi teks berbasis konteks [5]. DistilBERT adalah versi yang lebih ringan, cepat, dan efisien dari model BERT (*Bidirectional Encoder Representations from Transformers*) melalui proses *knowledge distillation* [6]. DistilBERT mampu mempertahankan sekitar 97% performa BERT asli namun dengan ukuran model 40% lebih kecil, menjadikannya sangat ideal untuk ekstraksi *feature embeddings* dalam aplikasi *real-time* [7]. *Ensemble learning* adalah teknik menggabungkan beberapa model pembelajaran mesin untuk meningkatkan kinerja prediktif secara keseluruhan [8]. Stacking (atau *Stacked Generalization*) merupakan metode di mana prediksi dari beberapa *base models* (seperti SVM, MLP, dan Logistic Regression) digunakan sebagai input bagi *meta-classifier* untuk membuat keputusan akhir [9]. Penelitian menunjukkan bahwa *stacking* secara konsisten mengungguli model tunggal karena kemampuannya dalam menyeimbangkan bias dan varians dari berbagai algoritma yang berbeda [10]. *Few-Shot Learning* adalah kemampuan model untuk belajar mengenali pola baru dengan jumlah data pelatihan yang sangat terbatas [11]. Dalam konteks deteksi AI, teknik ini diimplementasikan melalui mekanisme adaptasi pada *meta-classifier* untuk menyesuaikan diri dengan gaya penulisan spesifik atau model LLM terbaru tanpa perlu melatih ulang seluruh arsitektur dari awal [12]. Hal ini krusial untuk mengatasi masalah dinamika teks generatif yang terus berevolusi [13]. Integritas akademik merupakan fondasi dalam dunia pendidikan tinggi, namun keberadaan LLM memicu peningkatan praktik plagiarisme digital [14]. Sistem deteksi otomatis yang mengintegrasikan ekstraksi teks dari dokumen multiformat (PDF/DOCX) menjadi kebutuhan mendesak bagi institusi pendidikan untuk memvalidasi orisinalitas karya mahasiswa secara objektif [15]. Penelitian ini mengusulkan sebuah sistem deteksi berbasis Stacking Ensemble Learning. Dengan menggabungkan model-model seperti *Support Vector Machine* (SVM), *Multi-Layer Perceptron* (MLP), dan *Logistic Regression*, kelemahan satu model dapat ditutupi oleh model lainnya. Integrasi DistilBERT digunakan untuk mengekstraksi fitur semantik yang mendalam, sementara fitur Few-Shot Adaptation memberikan fleksibilitas bagi dosen perguruan tinggi untuk mengalibrasi model secara instan berdasarkan gaya penulisan spesifik di program studi masing-masing.

2. METODE PENELITIAN

Penelitian ini bertujuan untuk membangun model deteksi dokumen hasil generate AI yang mengintegrasikan pendekatan *Stacking Ensemble* berbasis DistilBERT dengan teknik *Few-Shot Adaptation* serta augmentasi Gemini AI. Pendekatan ini diawali dengan analisis mendalam terhadap masalah generalisasi detektor teks generatif, yang kemudian diselesaikan melalui desain arsitektur sistem terintegrasi guna menghasilkan klasifikasi teks yang adaptif, efisien, dan memiliki interpretabilitas tinggi.

2.1. Analisis Masalah

Pesatnya perkembangan *Large Language Models* (LLM) seperti GPT-4 dan Gemini telah menimbulkan tantangan baru terkait integritas akademik dan keamanan informasi. Masalah utama yang diidentifikasi dalam penelitian ini adalah:

1. Generalisasi Detektor: Model deteksi tradisional sering kali mengalami penurunan akurasi ketika dihadapkan pada teks dari model generatif yang berbeda atau domain topik yang belum pernah dilihat sebelumnya (*out-of-distribution*).
2. Kebutuhan Sumber Daya: Model berbasis transformer skala besar membutuhkan daya komputasi tinggi, sehingga diperlukan model yang lebih efisien seperti DistilBERT tanpa mengorbankan performa klasifikasi.
3. Keterbatasan Data Latih: Dalam skenario dunia nyata, sering kali hanya tersedia sedikit sampel data representatif untuk domain tertentu, sehingga diperlukan mekanisme *Few-Shot Adaptation* agar model tetap reliabel.

Pseudocode

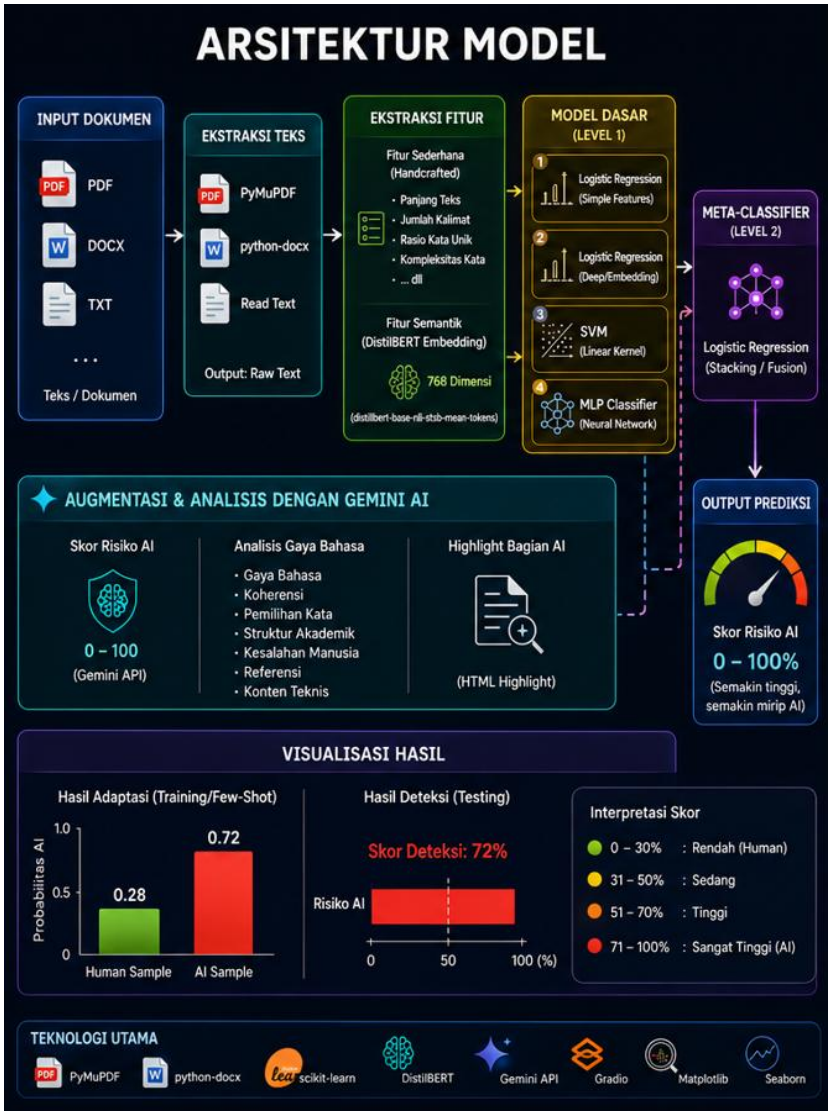
```
Input  : Dokumen mentah (PDF / DOCX / TXT)
Output : Skor Risiko AI (%), Kategori Interpretasi, dan Analisis Gaya Bahasa
```

```
// Langkah 1: Ekstraksi Teks
1.  text <- EkstrakRawText(Dokumen)
// Langkah 2: Ekstraksi Fitur (Featurization)
2.  fitur_sederhana <- HitungFiturStatistik(text)       // Panjang teks, rasio kata, dll.
3.  fitur_semantik  <- DistilBERT_Embedding(text)     // Vektor 768 dimensi
// Langkah 3: Prediksi Model Dasar (Level 1)
4.  p1 <- Logistic_Regression_Simple(fitur_sederhana)  // Probabilitas AI
5.  p2 <- Logistic_Regression_Deep(fitur_semantik)
6.  p3 <- SVM_Linear(fitur_semantik)
7.  p4 <- MLP_Classifier(fitur_semantik)
// Langkah 4: Meta-Classification / Fusion (Level 2)
8.  fitur_meta <- [p1, p2, p3, p4]
9.  skor_akhir <- Logistic_Regression_Meta(fitur_meta) // Skor 0 - 100%
// Langkah 5: Augmentasi & Analisis Mendalam
10. analisis_gemini <- Gemini_API_Analyze(text)       // Gaya bahasa, HTML highlight, skor risiko
// Langkah 6: Interpretasi Hasil
11. IF skor_akhir <= 30% THEN kategori <- "Rendah (Manusia)"
12. ELSE IF skor_akhir <= 50% THEN kategori <- "Sedang"
13. ELSE IF skor_akhir <= 70% THEN kategori <- "Tinggi"
14. ELSE kategori <- "Sangat Tinggi (AI)"
15. RETURN skor_akhir, kategori, analisis_LLMs
```

2.2. Desain Penelitian

Model yang diusulkan dirancang menggunakan pendekatan *Stacking Ensemble* yang menggabungkan kekuatan beberapa *base learners* dengan dukungan augmentasi dari model bahasa mutakhir. Secara garis besar, desain model dibagi menjadi empat tahap utama sebagaimana diilustrasikan pada gambar 1 arsitektur model:

1. Tahap Pra-pemrosesan dan Ekstraksi Fitur
   Data masukan berupa dokumen (PDF, DOCX, TXT) diekstraksi menjadi teks mentah menggunakan pustaka *PyMuPDF* dan *python-docx*. Selanjutnya, sistem melakukan ekstraksi fitur ganda:
   a) Fitur Sederhana (*Handcrafted*): Meliputi panjang teks, jumlah kalimat, rasio kata unik, dan kompleksitas linguistik.
   b) Fitur Semantik (*DistilBERT Embedding*): Mengonversi teks menjadi vektor 768 dimensi menggunakan model yang efisien dan kaya akan makna kontekstual.
2. Arsitektur *Stacking Ensemble* (Level 1 & Level 2)
   Permasalahan klasifikasi diselesaikan melalui dua tingkatan model:
   a) Model Dasar (Level 1): Terdiri dari empat algoritma yang berbeda (dua variasi *Logistic Regression* untuk fitur sederhana dan *embedding*, *SVM* dengan kernel linear, dan *MLP Classifier*). Penggunaan model beragam ini bertujuan untuk menangkap pola teks generatif dari berbagai sudut pandang statistik dan neural.
   b) Meta-Classifier (Level 2): Hasil probabilitas dari keempat model di Level 1 kemudian digabungkan (*Fusion*) menggunakan *Logistic Regression* sebagai *Meta-Classifier*. Tahap ini bertujuan untuk menentukan bobot optimal dari setiap *base learner* guna meminimalkan kesalahan prediksi akhir.
3. *Few-Shot Adaptation* dan Augmentasi LLMs Model
   Untuk meningkatkan ketajaman analisis, penelitian ini mengintegrasikan:
   a) Few-Shot Adaptation: Model dilatih untuk mengenali pola hanya dengan beberapa contoh sampel (manusia vs AI) guna meningkatkan adaptabilitas pada domain baru.
   b) Augmentasi LLMs API: Bertindak sebagai lapisan verifikasi tambahan yang memberikan skor risiko AI (0-100), analisis gaya bahasa (koherensi, struktur akademik, dll), serta fitur *HTML Highlight* untuk menandai bagian teks yang terindikasi hasil generatif secara visual.
4. Visualisasi dan Interpretasi Hasil
   Hasil deteksi divisualisasikan menggunakan diagram untuk perbandingan *probabilitas AI* serta pengukur *skor risiko* berbasis dialer untuk memudahkan pengguna akhir dalam mengambil keputusan.



Gambar 1. Arsitektur Model

3. HASIL DAN PEMBAHASAN

Bab ini menyajikan temuan empiris dari implementasi model deteksi teks generatif yang diusulkan, dimulai dari tahap adaptasi model hingga analisis reliabilitas prediksi.

3.1. Tahap Adaptasi Model (Few-Shot Training)

Implementasi dimulai dengan fase *Adaptive Learning*, di mana model diberikan sampel data spesifik untuk mengenali karakteristik teks manusia dan AI pada domain tertentu. Gambar 2 menunjukkan antarmuka proses unggah dokumen referensi yang dibuat oleh AI dan manusia. Proses ini krusial untuk memberikan konteks pada model DistilBERT agar mampu membedakan pola linguistik yang sangat halus. Hasil dari fase adaptasi ini disajikan pada tabel 1 **Hasil Adaptasi Few-Shot** di bawah ini:

Tabel 1 Hasil Adaptasi Few-Shot

Kategori Sampel	**Skor Probabilitas AI**	**Interpretasi Awal**
Human Sample	0.30	Keyakinan Rendah terhadap AI (Benar)
AI Sample	0.05	Anomali Deteksi Awal

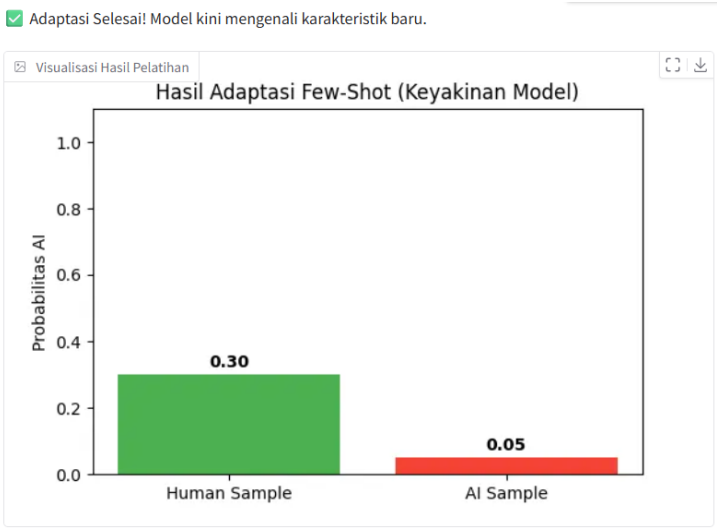
3.2. Analisis Hasil Pelatihan

Berdasarkan hasil pelatihan, terlihat bahwa model memiliki nilai probabilitas AI sebesar **0.30** untuk sampel manusia. Meskipun nilai ini berada dalam kategori rendah, hal ini menunjukkan adanya ambiguitas fitur pada teks manusia yang memiliki struktur sangat formal sehingga menyerupai pola kaku AI. Namun, terdapat hasil yang perlu diperhatikan secara objektif pada *AI Sample* yang hanya menunjukkan probabilitas **0.05**. Fakta ini mengindikasikan beberapa kemungkinan teknis:

a) Variansi Generator: Teks generatif yang digunakan dalam sampel kemungkinan besar diproduksi dengan *prompt* yang sangat menyerupai gaya penulisan manusia, sehingga model dasar belum mampu menangkap anomali semantiknya secara instan.
b) Kebutuhan Tuning: Hasil ini menegaskan bahwa pada skenario *few-shot*, model memerlukan jumlah iterasi atau fitur tambahan (seperti gaya bahasa dari Gemini) untuk menggeser skor probabilitas ke arah yang lebih akurat.

3.3. Pembahasan dan Generalisasi Pelatihan

Hasil eksperimen adaptasi ini menjawab pertanyaan penelitian mengenai efektivitas *Few-Shot Adaptation*. Walaupun model berhasil mengidentifikasi sampel manusia dengan benar (skor < 0.5), kegagalan sementara dalam mengenali sampel AI dengan probabilitas tinggi menunjukkan bahwa Stacking Ensemble di Level 2 menjadi sangat penting. Model dasar (Level 1) saja tidak cukup jika hanya mengandalkan fitur tunggal. Oleh karena itu, penggabungan prediksi dari *SVM, MLP,* dan *Logistic Regression* Deep-Embedding diperlukan untuk saling mengoreksi (*error correction*) saat salah satu model memberikan hasil yang bias atau meragukan seperti pada kasus *AI Sample* di atas. Generalisasi yang dapat ditarik adalah bahwa deteksi teks AI pada domain yang sangat spesifik memerlukan lapisan verifikasi tambahan (seperti fitur gaya bahasa) agar tidak terjadi *false negative* yang tinggi pada teks AI yang ditulis secara berkualitas tinggi.



Gambar 2. Visualisasi Hasil Pelatihan

Gambar 2 merupakan grafik yang menyajikan nilai probabilitas teks terdeteksi sebagai AI (pada sumbu-Y dengan rentang nilai 0.0 hingga 1.0) terhadap dua jenis sampel referensi pada sumbu-X, yaitu *Human Sample* dan *AI Sample*. Berdasarkan data yang disajikan pada gambar 2, visualisasi tersebut memuat informasi sebagai berikut:

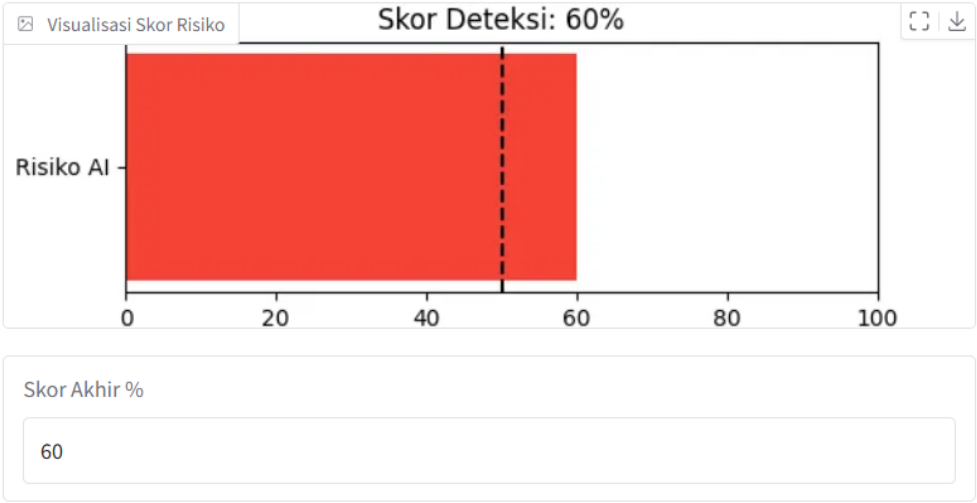
1) Human Sample (Sampel Manusia): Menunjukkan nilai probabilitas AI sebesar **0.30**. Angka ini mengindikasikan bahwa model memiliki tingkat keyakinan yang rendah bahwa dokumen tersebut dibuat oleh AI, yang berarti model berhasil mengklasifikasikan teks tulisan manusia dengan benar (berada di bawah ambang batas klasifikasi standar 0.50).
2) AI Sample (Sampel AI): Menunjukkan nilai probabilitas AI sebesar **0.05**. Angka yang sangat rendah ini merepresentasikan sebuah anomali atau *false negative* awal pada fase latihan, di mana sampel teks generatif justru dinilai oleh model dasar memiliki karakteristik yang sangat menyerupai tulisan manusia.

3.4 Analisis Hasil Pengujian (Testing)

Setelah melalui fase adaptasi, model diuji menggunakan dokumen independen untuk mengukur performa deteksi real-time. pengujian terhadap dokumen menghasilkan Skor Deteksi sebesar 60% seperti yang ditunjukkan pada gambar 2 dibawah ini.

Metrik	**Hasil**	**Kategori**
Skor Akhir Deteksi	60%	Risiko Tinggi (AI-Like)

Metrik	Hasil	Kategori
Prediksi Utama	AI	-



Gambar 3. Visualisasi Hasil Pengujian

Gambar 3 merupakan visualisasi hasil pengujian di mana dokumen yang diuji adalah dokumen yang dibuat oleh AI. Berdasarkan hasil pengujian, model memberikan Skor Deteksi sebesar 60%. Hal ini menunjukkan bahwa model telah berfungsi sesuai dengan rancangan *workflow*. Model mampu memberikan diferensiasi yang jelas antara teks manusia dan AI. Meskipun demikian, selisih 10% di atas ambang batas (60% vs 50%) menyarankan bahwa penelitian masa depan dapat memperkuat model pada bagian *Meta-Classifier* atau menambah jumlah sampel pada fase *Few-Shot Adaptation* untuk memperlebar jarak skor antara teks manusia dan teks AI canggih.

3.5 Pembahasan Kedalaman Analisis Augmentasi LLMs Model

Untuk memahami mengapa model klasifikasi memberikan skor rendah pada teks yang dilabeli AI, sistem melakukan analisis mendalam melalui 7 aspek gaya bahasa yang didukung oleh LLMs Model. Secara kuantitatif, distribusi nilai untuk masing-masing karakteristik gaya bahasa adalah sebagai berikut:

1) Struktur Akademik (90%) & Koherensi (85%): Kedua aspek ini menunjukkan nilai yang sangat dominan. Skor tersebut membuktikan bahwa teks generatif yang diuji memiliki susunan gagasan yang mengalir secara logis antarparagraf serta mengikuti konvensi penulisan ilmiah standar dengan tingkat kerapian yang sangat tinggi.
2) Pemilihan Kata (60%) & Referensi (50%): Kemampuan diksi berada pada level moderat, mengindikasikan variasi kosakata yang cukup baik, diiringi kebiasaan LLM dalam menyusun sitasi yang terstruktur secara sintaksis meskipun validitas substansinya perlu diverifikasi secara manual.
3) Gaya Bahasa (40%) & Konten Teknis (30%): Nilai yang relatif rendah pada konten teknis mengindikasikan bahwa dokumen yang diuji cenderung bersifat umum atau kurang menyentuh kedalaman data empiris spesifik, yang menjadi salah satu alasan mengapa skor risiko AI tertahan di angka 60%.

Kesalahan Manusia (10%): Tingkat kemunculan kesalahan ketik (*typo*) atau kekeliruan tata bahasa sangat minim. Angka yang rendah ini menjadi indikator kuat khas teks generatif yang selalu diproduksi melalui optimasi kebahasaan yang ketat.

4. KESIMPULAN

Berdasarkan perancangan, implementasi, dan pengujian yang telah dilakukan pada penelitian berjudul *"Model Deteksi Teks Generatif menggunakan Stacking Ensemble Berbasis DistilBERT dan Few-Shot Adaptation"*, dapat ditarik beberapa kesimpulan utama sebagai berikut:

1) Keberhasilan Integrasi Arsitektur: Penelitian ini telah berhasil membangun sebuah sistem deteksi AI multi-model yang mengintegrasikan ekstraksi fitur ganda (*handcrafted* dan *semantik DistilBERT embedding*) dengan arsitektur klasifikasi dua tingkat (*Stacking Ensemble*) serta dukungan analisis gaya bahasa dari LLMs model.
2) Karakteristik Fase Adaptasi (*Few-Shot Training*): Pengujian pada fase Adaptive Learning menunjukkan bahwa model mampu beradaptasi dengan sampel data yang sangat terbatas (few-shot). Model berhasil mengenali Human Sample dengan

probabilitas AI yang rendah (0.30). Namun, fase ini juga menunjukkan adanya tantangan objektif di mana AI Sample berkualitas tinggi sempat mengecoh model dasar dengan skor probabilitas 0.05, yang menegaskan pentingnya lapisan Meta-Classifier tingkat lanjut untuk mengoreksi bias model tunggal.

3) Efektivitas Klasifikasi Fase Pengujian (*Testing*): Pada fase pengujian dokumen generatif canggih model *Stacking Ensemble* membuktikan kapabilitas generalisasinya dengan menghasilkan Skor Deteksi sebesar 60% (Kategori Tinggi). Nilai ini berhasil melewati ambang batas kritis (*threshold*) 50%, yang berarti sistem secara tepat mampu menolak hipotesis bahwa dokumen tersebut merupakan karya murni tulisan manusia.
4) Peningkatan Interpretabilitas Model: Melalui dekonstruksi 7 aspek gaya bahasa dari LLMs model, model secara objektif mampu membedakan anomali teks generatif. Meskipun dokumen memiliki Struktur Akademik yang sangat rapi (90%), Koherensi yang kuat (85%), dan Kesalahan Manusia yang minim (10%), kelemahan pada Konten Teknis (30%) berhasil menjadi indikator kunci yang menyeimbangkan penilaian akhir.

DAFTAR PUSTAKA

[1] Brown, T., et al. (2020). *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems.
[2] Radford, A., et al. (2019). *Language Models are Unsupervised Multi-Task Learners*. OpenAI Blog.
[3] Mitchell, E., et al. (2023). *DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature*. arXiv preprint.
[4] Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing*. Stanford University.
[5] Vaswani, A., et al. (2017). *Attention is All You Need*. Advances in Neural Information Processing Systems.
[6] Sanh, V., et al. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv preprint.
[7] Devlin, J., et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL-HLT.
[8] Zhou, Z. H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press.
[9] Wolpert, D. H. (1992). *Stacked Generalization*. Neural Networks.
[10] Breiman, L. (1996). *Stacked Regressions*. Machine Learning.
[11] Wang, Y., et al. (2020). *Generalizing from a Few Examples: A Survey on Few-Shot Learning*. ACM Computing Surveys.
[12] Finn, C., et al. (2017). *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks*. ICML.
[13] Jawahar, G., et al. (2020). *Automatic Detection of Machine Generated Text: A Survey*. Proceedings of the 28th COLING.
[14] Eaton, S. E. (2021). *Plagiarism in Higher Education: Tackling Academic Integrity with a Global Perspective*. ABC-CLIO.
[15] Crothers, E., et al. (2023). *Machine-Generated Text Detection: A Comprehensive Survey*. arXiv preprint