JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i3.9539

PERBANDINGAN AKURASI MODEL AI DALAM MENILAI KEPRIBADIAN PENGGUNA MENGGUNAKAN PENDEKATAN NLP (NATURAL LANGUAGE PROCESSING)

Ryan Hadi Ardhiansyah[1*], Rahmat Budiarsa[2]

[1,2]Universitas Esa Unggul; Jl. Arjuna Utara No.9, Duri Kepa, Kec. Kb. Jeruk, Kota Jakarta Barat, Daerah Khusus Ibukota Jakarta 11510; (021) 39529950

Keywords:
Kecerdasan Buatan;
Pemrosesan Bahasa Alami;
Lima Dimensi Kepribadian Utama;
Pembelajaran Mendalam;
Model Berbasis Aturan.

Corespondent Email:
ryanhadi412003@student.esaunggul.ac.id

Abstrak. Penilaian kepribadian berbasis teks merupakan salah satu aplikasi NLP yang berkembang pesat, namun pemilihan model yang tepat antara efisiensi dan akurasi masih menjadi tantangan. Penelitian ini membandingkan efektivitas pendekatan deep learning (BERT dan DistilBERT) dengan metode rule-based dalam memprediksi lima dimensi kepribadian Big Five dari data teks. Kedua model transformer dioptimasi menggunakan Task-Specific Head Tuning dikombinasikan dengan kalibrasi threshold, dan dievaluasi pada dua kategori panjang teks: teks pendek (≤128 token) dan teks panjang (≥256 token) dari dataset agregat sebanyak 37.630 sampel. Hasil menunjukkan bahwa DistilBERT dengan full fine-tuning mencapai performa terbaik pada teks pendek (F1-Score: 0,6751; nMCC: 0,6807), sementara BERT dengan full fine-tuning unggul pada teks panjang (nMCC: 0,6258). Metode rule-based secara konsisten menghasilkan performa lebih rendah dengan nMCC tertinggi hanya 0,5288. Analisis per dimensi menunjukkan bahwa Openness paling mudah dideteksi pada teks pendek (F1-Score: 0,8405), sedangkan Agreeableness membutuhkan konteks lebih panjang (F1-Score: 0,7198). Penelitian ini menyimpulkan bahwa DistilBERT merupakan pilihan optimal untuk lingkungan dengan sumber daya komputasi terbatas, sementara BERT tetap unggul untuk analisis teks panjang yang kompleks.

Abstract. *Text-based personality assessment is a rapidly growing NLP application, yet selecting the appropriate model that balances efficiency and accuracy remains challenging. This study compares the effectiveness of deep learning approaches (BERT and DistilBERT) against a rule-based method in predicting Big Five personality traits from text data. Both transformer models were optimized using Task-Specific Head Tuning combined with threshold calibration, and evaluated on two text length categories: short texts (≤128 tokens) and long texts (≥256 tokens) from an aggregated dataset of 37,630 samples. Results indicate that DistilBERT with full fine-tuning achieved the best performance on short texts (F1-Score: 0.6751; nMCC: 0.6807), while BERT with full fine-tuning outperformed on long texts (nMCC: 0.6258). Rule-based methods consistently underperformed with the highest nMCC reaching only 0.5288. Per-dimension analysis revealed that Openness was most easily detected on short texts (F1-Score: 0.8405), whereas Agreeableness required longer context (F1-Score: 0.7198). This study concludes that DistilBERT is*

the optimal choice for resource-constrained environments, while BERT remains superior for analyzing long and contextually complex texts.

1. PENDAHULUAN

Ketergantungan yang semakin meningkat terhadap kecerdasan buatan (AI) telah mengubah cara sistem digital menginterpretasikan dan beradaptasi terhadap perilaku manusia. Dalam domain natural language processing (NLP), salah satu aplikasi yang paling menonjol adalah penilaian kepribadian secara komputasional, di mana sifat kepribadian pengguna disimpulkan langsung dari data teks [1]. Kemampuan untuk mengenali kepribadian dari isyarat linguistik memfasilitasi pengembangan interaksi yang lebih personal dalam domain seperti konseling digital, pemasaran psikografis, dan analitik media sosial [2]. Meskipun memiliki potensi besar, pencapaian akurasi prediksi yang tinggi tetap menjadi tantangan utama, karena model yang ada bervariasi secara signifikan dalam kemampuannya menangkap sifat psikologis yang bernuansa dari data teks [3].

Model deep learning, khususnya arsitektur berbasis transformer seperti BERT dan DistilBERT, telah menunjukkan performa yang kuat dalam tugas klasifikasi teks dan pembelajaran representasi [4]. Pemahaman kontekstual bidireksional yang dimilikinya memungkinkan model-model ini mengungguli pendekatan tradisional dalam berbagai tolok ukur NLP [5]. Namun, model-model ini membutuhkan sumber daya komputasi yang besar sehingga kurang praktis untuk aplikasi ringan [6]. Sebaliknya, metode rule-based menawarkan interpretabilitas yang lebih tinggi, tetapi umumnya kurang akurat dibandingkan deep learning [7]. Pertimbangan ini memunculkan pertanyaan kritis mengenai keseimbangan antara performa dan efisiensi dalam penerapan sistem prediksi kepribadian di dunia nyata.

Tantangan lain yang teridentifikasi adalah sensitivitas model NLP terhadap panjang teks. Studi-studi sebelumnya menunjukkan bahwa input teks yang lebih panjang memberikan sinyal linguistik yang lebih kaya sehingga berpotensi meningkatkan akurasi prediksi, sementara teks pendek dapat membatasi pemahaman kontekstual [8]. Namun, inkonsistensi performa lintas panjang teks masih kurang dieksplorasi dalam konteks prediksi kepribadian [9]. Mengatasi kesenjangan ini memerlukan evaluasi sistematis mengenai bagaimana arsitektur model berperilaku pada panjang input yang berbeda.

Berdasarkan latar belakang tersebut, penelitian ini memiliki tiga kontribusi utama. Pertama, penelitian ini menyediakan evaluasi komparatif model deep learning (BERT, DistilBERT) terhadap baseline berbasis aturan untuk memprediksi sifat kepribadian Big Five. Kedua, penelitian ini mengkaji pengaruh panjang teks dengan mengkategorikan dataset menjadi dua kelompok, yaitu teks pendek (≤128 token) dan teks panjang (≥256 token). Ketiga, penelitian ini memperkenalkan strategi optimasi efisien melalui Task-Specific Head Tuning yang dikombinasikan dengan kalibrasi threshold, sebagai alternatif yang ringan namun efektif dibandingkan full fine-tuning. Secara keseluruhan, kontribusi-kontribusi ini menyoroti keunggulan dan pertimbangan dari berbagai strategi pemodelan, serta memberikan panduan praktis untuk penilaian kepribadian komputasional dalam aplikasi berbasis NLP.

2. TINJAUAN PUSTAKA

2.1. Pemrosesan Bahasa Alami (Natural Language Processing/NLP)

Pemrosesan Bahasa Alami (Natural Language Processing atau NLP) adalah cabang dari AI yang berfokus pada interaksi antara komputer dan bahasa manusia. NLP bertujuan untuk memungkinkan komputer memahami, menginterpretasikan, dan menghasilkan bahasa yang alami bagi manusia. NLP berkembang dari sistem berbasis aturan di awal perkembangannya menuju penerapan model pembelajaran mendalam [10]. Dalam aplikasi psikologi, NLP digunakan untuk mengungkapkan karakteristik psikologis seseorang berdasarkan data teks, seperti unggahan media sosial atau hasil survei, serta telah memberikan alat yang kuat bagi psikologi komputasional untuk menganalisis teks dalam skala besar [11].

2.2. Model Deep Learning

Deep learning merupakan bagian dari machine learning yang menggunakan jaringan saraf tiruan berlapis-lapis (deep neural networks) untuk mempelajari representasi data yang kompleks [12]. Dalam penelitian ini, model deep learning yang menjadi fokus utama adalah arsitektur Transformer, khususnya BERT dan DistilBERT.

BERT (Bidirectional Encoder Representations from Transformers) adalah model bahasa berbasis arsitektur Transformer yang dilatih pada korpus teks berukuran besar menggunakan tugas Masked Language Modeling (MLM) dan Next Sentence Prediction (NSP). Pelatihan ini memungkinkan BERT memahami konteks bahasa secara mendalam dari kedua arah sehingga menghasilkan representasi kata yang kaya makna [4]. Efektivitas BERT telah terbukti dalam banyak penerapan, di mana proses fine-tuning untuk tugas spesifik seperti prediksi kepribadian mampu mencapai kinerja state-of-the-art [3].

DistilBERT merupakan versi ringan dan lebih efisien dari BERT yang diperoleh melalui teknik knowledge distillation [6]. DistilBERT memiliki arsitektur yang lebih kecil dan performa inferensi yang lebih cepat dibandingkan BERT, namun tetap mempertahankan sebagian besar kemampuan pemahaman bahasa dari model induknya. Penggunaan DistilBERT dalam penelitian ini bertujuan untuk mengeksplorasi trade-off antara akurasi dan efisiensi komputasi dalam prediksi kepribadian berbasis teks [6].

2.3. Model Rule-Based

Model rule-based dalam NLP menggunakan aturan-aturan linguistik yang telah ditentukan sebelumnya untuk menganalisis teks dan mengekstrak informasi [5], [7]. Aturan-aturan ini dapat berupa pola leksikal, sintaksis, atau semantik yang berkorelasi dengan sifat-sifat kepribadian. Meskipun mudah diinterpretasi, model rule-based umumnya kesulitan dalam menangani variasi dan ambiguitas bahasa alami sehingga menghasilkan akurasi yang lebih rendah dibandingkan pendekatan deep learning [7].

2.4. Kepribadian Big Five (OCEAN)

Kepribadian Big Five merupakan model yang banyak digunakan dan diterima secara luas dalam psikologi kepribadian. Model ini mengklasifikasikan kepribadian menjadi lima dimensi utama, yaitu Openness (Keterbukaan), Conscientiousness (Ketelitian), Extraversion (Ekstroversi), Agreeableness (Keramahan), dan Neuroticism (Neurotisisme) [13]. Setiap dimensi mewakili spektrum karakteristik kepribadian, misalnya Extraversion menggambarkan seseorang dari yang cenderung pendiam (introvert) hingga yang ramah dan mudah bergaul (extrovert). Pemilihan kerangka Big Five ini juga didukung oleh penelitian sebelumnya yang menunjukkan bahwa model ini lebih informatif dan stabil untuk tugas prediksi kepribadian berbasis teks dibandingkan alternatif lain seperti MBTI [7].

2.5. Metrik Evaluasi

Evaluasi performa model dalam penelitian ini menggunakan beberapa metrik standar klasifikasi. Akurasi mengukur persentase keseluruhan prediksi yang benar. Presisi mengukur ketepatan prediksi positif, sementara recall mengukur kemampuan model mengidentifikasi seluruh kasus positif yang relevan. F1-Score dihitung sebagai rata-rata harmonik dari presisi dan recall untuk memberikan skor tunggal yang menyeimbangkan keduanya [14].

Untuk mengatasi keterbatasan metrik-metrik tersebut pada dataset dengan distribusi kelas yang tidak seimbang, Matthews Correlation Coefficient (MCC) digunakan sebagai metrik utama tambahan. MCC memperhitungkan seluruh komponen confusion matrix (true positives, true negatives, false positives, dan false negatives) sehingga memberikan ukuran korelasi yang lebih representatif antara prediksi dan label aktual [15].

3. METODE PENELITIAN

Penelitian ini mengadopsi desain penelitian komparatif kuantitatif untuk mengevaluasi performa pendekatan deep learning dan rule-based dalam tugas prediksi kepribadian secara komputasional. Tujuannya adalah untuk menentukan sejauh mana model berbasis Transformer, khususnya BERT dan DistilBERT, mengungguli baseline berbasis leksikon dalam memprediksi sifat kepribadian

Big Five dari teks [5]. Desain penelitian ini menekankan akurasi prediktif sekaligus efisiensi komputasi, sehingga memungkinkan perbandingan yang seimbang antar metodologi.

3.1. Dataset dan Praproses

3.1.1. Pengumpulan Dataset

Penelitian ini menggunakan dataset komposit yang diagregasi dari enam sumber yang tersedia secara publik untuk memastikan generalisabilitas temuan. Sumber-sumber tersebut mencakup teks dari berbagai domain seperti platform media sosial, esai pribadi, dan transkrip dialog, guna memastikan representasi gaya linguistik dan konteks yang beragam [16], [17], [18]. Sebelum agregasi, struktur setiap sumber distandarisasi ke format umum yang terdiri dari satu field teks dan lima label biner yang bersesuaian dengan sifat kepribadian Big Five (Openness, Conscientiousness, Extraversion, Agreeableness, dan Neuroticism). Setelah proses agregasi dan penyaringan entri yang tidak lengkap, dataset akhir terdiri dari 37.630 sampel teks.

3.1.2. Praproses Data

Pipeline praproses terstandarisasi diterapkan pada seluruh dataset untuk membersihkan dan menormalisasi teks, guna memastikan bahwa model belajar dari fitur linguistik yang bermakna. Pipeline ini mencakup beberapa tahap otomatis, yaitu: (a) penghapusan artefak non-linguistik seperti URL, tag HTML, sebutan pengguna, dan karakter khusus; (b) normalisasi teks yang meliputi konversi seluruh karakter ke huruf kecil dan standardisasi spasi; serta (c) verifikasi akhir untuk memastikan integritas data [5].

3.1.3. Pengacuan Pustaka. (Reference Library.)

Untuk menginvestigasi pengaruh panjang teks terhadap performa model, dataset yang telah diproses distratifikasi menjadi dua subset berdasarkan jumlah token [19]. Kedua subset didefinisikan sebagai berikut:

- Short Texts: Sampel yang mengandung 128 token atau kurang (N = 31.584).
- Long Texts: Sampel yang mengandung 256 token atau lebih (N = 6.046).

Selanjutnya, setiap subset distratifikasi dibagi secara independen dan acak menjadi set pelatihan (80%), set validasi (10%), dan set pengujian (10%). Set pelatihan digunakan secara eksklusif untuk melatih model, set validasi digunakan untuk penyetelan hyperparameter dan proses kalibrasi threshold, sementara set pengujian disimpan dan hanya digunakan sekali untuk evaluasi akhir yang tidak bias. Random seed tetap digunakan sepanjang proses pembagian untuk memastikan reproduktibilitas eksperimen.

3.2. Arsitektur Model

3.2.1. Model Deep Learning

Penelitian ini menggunakan dua model Transformer pre-trained dari pustaka Hugging Face, yaitu BERT (bert-base-uncased) dan DistilBERT (distilbert-base-uncased). Untuk menyeimbangkan kekuatan prediktif dengan efisiensi komputasi, penelitian ini mengimplementasikan strategi Parameter-Efficient Fine-Tuning (PEFT) melalui Task-Specific Head Tuning [20]. Pendekatan ini terdiri dari dua komponen utama:

- Frozen Backbone: Lapisan inti model Transformer pre-trained dibekukan sehingga parameternya tidak diperbarui selama pelatihan, dengan demikian pengetahuan linguistik yang telah dipelajari dari korpus berskala besar tetap terjaga [4], [6].
- Classification Head: Sebuah head klasifikasi ringan ditambahkan di atas backbone yang dibekukan. Head ini terdiri dari sebuah lapisan linear yang memetakan keluaran backbone ke dimensi tersembunyi (256), diikuti oleh fungsi aktivasi ReLU, lapisan dropout untuk regularisasi, dan lapisan linear akhir yang menghasilkan lima logit yang bersesuaian dengan sifat kepribadian Big Five.

Selama fase pelatihan, hanya parameter *classification head* yang diperbarui. Pendekatan ini terbukti efektif sebagai metode *parameter-efficient tuning* yang mampu mencapai performa kuat dengan mengadaptasi *head* khusus pada tugas target [21].

3.2.2. Model Rule-Based (Baseline)

Sebagai tolok ukur performa, model baseline berbasis aturan diimplementasikan menggunakan pendekatan berbasis leksikon yang beroperasi pada Lexical Hypothesis [22]. Model ini sepenuhnya bebas pelatihan

(training-free) dengan mekanisme sebagai berikut:

- Lexicon Matching: Token dari teks masukan dicocokkan dengan leksikon psikologis yang telah divalidasi, di mana kata-kata tertentu diasosiasikan dengan skor untuk kelima sifat Big Five.
- Score Aggregation: Skor akhir untuk setiap sifat dihitung dengan mengagregasi skor seluruh kata kunci yang cocok dalam teks yang diberikan.

3.3. Kerangka Evaluasi

3.3.1. Pelatihan dan Kalibrasi Model Deep Learning

Model deep learning dilatih dan dikalibrasi melalui proses dua tahap. Pada tahap pertama, fine-tuning dilakukan dengan hanya menyesuaikan parameter classification head sambil membekukan lapisan Transformer pre-trained. Pencarian hyperparameter sistematis dilakukan pada learning rate berkisar antara 1e-5 hingga 5e-5 dan batch size 16 dan 32. Pelatihan model menggunakan optimizer AdamW dikombinasikan dengan fungsi Binary Cross-Entropy with Logits Loss.

Pada tahap kedua, dilakukan proses kalibrasi threshold. Nilai threshold optimal ditentukan secara independen untuk setiap dimensi kepribadian dengan mengevaluasi nilai threshold dalam rentang 0,0 hingga 1,0 pada set validasi, untuk memaksimalkan skor komposit dari F1-Score dan normalized Matthews Correlation Coefficient (MCC). Prosedur kalibrasi ini berperan penting dalam menyeimbangkan presisi dan recall sehingga meningkatkan keandalan prediksi akhir [23].

3.3.2. Kalibrasi dan Evaluasi Model Rule-Based

Model *rule-based* tidak menjalani fase pelatihan. Namun, untuk memastikan perbandingan yang adil, prosedur kalibrasi *threshold* yang sama diterapkan pada skor leksikon mentah yang dihasilkan model pada set validasi. Penerapan metode kalibrasi berbasis data yang identik pada semua model memastikan bahwa perbandingan performa pada set pengujian merupakan cerminan langsung dari kemampuan representasi model, bukan artefak dari perbedaan batas keputusan.

3.4. Metrik Evaluasi

Performa setiap model dievaluasi secara kuantitatif menggunakan serangkaian metrik klasifikasi standar. Evaluasi mencakup Akurasi yang mengukur proporsi keseluruhan prediksi yang benar. Untuk penilaian yang lebih mendalam, Presisi, Recall, dan F1-Score juga digunakan. Presisi mengukur proporsi prediksi positif yang benar di antara seluruh prediksi positif, sementara Recall mengukur proporsi positif sejati yang berhasil diidentifikasi. F1-Score merupakan rata-rata harmonik dari Presisi dan Recall yang memberikan ukuran performa model yang seimbang [14].

Untuk memperhitungkan potensi ketidakseimbangan kelas dalam data sifat kepribadian, Matthews Correlation Coefficient (MCC) disertakan sebagai metrik evaluasi utama. MCC mencakup keempat kuadran confusion matrix sehingga menjadikannya indikator kualitas klasifikasi yang lebih andal dan seimbang [15]. Seluruh nilai MCC yang dilaporkan dinormalisasi (nMCC) menggunakan formula berikut:

$$\text{nMCC} = \frac{(MCC + 1)}{2}$$

Pada skala ternormalisasi ini, skor 1,0 merepresentasikan prediksi sempurna, 0,5 setara dengan performa acak, dan 0 mengindikasikan prediksi invers total.

4. HASIL DAN PEMBAHASAN

4.1. Performa Keseluruhan Model: Deep Learning vs. Rule-Based

Performa komprehensif setiap model dan strategi optimasi pada kedua skenario teks pendek dan teks panjang dirangkum dalam Tabel 1. Metrik yang dilaporkan merupakan rata-rata mikro keseluruhan (overall micro-averages) yang memberikan gambaran holistik kemampuan prediktif setiap model di seluruh lima dimensi kepribadian. Hasil yang disajikan berasal dari konfigurasi hyperparameter terbaik untuk setiap skenario spesifik.

Tabel 1. Perbandingan Performa Komprehensif Semua Model

Model & Metode Tuning	*Panjang Teks*	*Accuracy*	*F1-Score*	*nMCC*
	Pendek	0.5567	0.6289	0.6102

Model & Metode Tuning	*Panjang Teks*	*Accuracy*	*F1-Score*	*nMCC*
BERT (Head Tuning)	Panjang	0.5736	0.6436	0.5926
BERT (Full Fine-Tuning)	Pendek	0.6313	0.6600	0.6615
	Panjang	0.5921	**0.6754**	**0.6258**
DistilBERT (Head Tuning)	Pendek	0.5822	0.6367	0.6259
	Panjang	0.5693	0.6639	0.6041
DistilBERT (Full Fine-Tuning)	**Pendek**	**0.6546**	**0.6751**	**0.6807**
	Panjang	**0.5974**	0.6694	0.6231
Rule-Based (Baseline)	Pendek	0.4444	0.6043	0.5288
	Panjang	0.5157	0.6415	0.5483

Catatan: Performa terbaik untuk setiap kategori panjang teks dicetak tebal.

Data yang disajikan pada Tabel 1 mengungkapkan beberapa temuan mendasar. Temuan utama adalah keunggulan yang konsisten dan signifikan dari model deep learning dibandingkan baseline rule-based. Pada skenario teks pendek, skor nMCC baseline sebesar 0,5288 jauh lebih rendah dibandingkan pendekatan deep learning paling ringan sekalipun, yaitu DistilBERT dengan head tuning (0,6259). Kesenjangan performa ini sejalan dengan eksplorasi algoritma machine learning secara lebih luas yang umumnya menemukan bahwa model yang mampu mempelajari representasi data kompleks mengungguli metode yang lebih sederhana [24]. Hal ini memvalidasi kebutuhan akan pemahaman kontekstual yang disediakan oleh arsitektur Transformer untuk prediksi kepribadian yang andal [5].

Kedua, hasil menunjukkan efektivitas yang jelas dari strategi optimasi full fine-tuning. Pada BERT maupun DistilBERT, full fine-tuning secara konsisten menghasilkan performa lebih tinggi dalam F1-Score dan nMCC dibandingkan pendekatan head tuning yang lebih terkekang. Sebagai contoh, pada teks pendek, skor nMCC DistilBERT meningkat dari 0,6259 (Head) menjadi 0,6807 (Full), sebuah peningkatan substansial dalam keandalan prediktif. Hal ini selaras dengan ekspektasi teoretis bahwa membiarkan seluruh model beradaptasi terhadap domain target memungkinkan pemahaman yang lebih mendalam terhadap pola linguistik spesifik tugas [4].

Terakhir, tabel ini menyoroti interaksi yang menarik antara arsitektur model dan panjang teks. DistilBERT dengan full fine-tuning muncul sebagai model berperforma terbaik pada teks pendek dengan skor tertinggi di seluruh metrik utama. Sebaliknya, BERT dengan full fine-tuning mengamankan skor nMCC tertinggi pada teks panjang, mengindikasikan bahwa arsitekturnya yang lebih besar mungkin lebih cocok untuk memanfaatkan informasi kontekstual yang lebih kaya.

4.2. Pengaruh Panjang Teks terhadap Performa

Hipotesis sentral penelitian ini adalah bahwa panjang teks secara signifikan memodulasi performa prediktif model penilaian kepribadian. Dengan membandingkan hasil dari dataset teks pendek (≤128 token) dan teks panjang (≥256 token), sebuah tren yang jelas dan konsisten muncul: peningkatan panjang teks pada umumnya berkorelasi dengan peningkatan performa model, meskipun dengan efek yang berbeda-beda di setiap dimensi kepribadian.

Peningkatan performa secara umum terlihat pada sebagian besar model sebagaimana ditunjukkan pada Tabel 1. Sebagai contoh, F1-Score baseline rule-based meningkat dari 0,6043 pada teks pendek menjadi 0,6415 pada teks panjang. Tren ini mengindikasikan bahwa teks yang lebih panjang memberikan volume bukti linguistik yang lebih besar sehingga memungkinkan model membuat prediksi yang lebih akurat. Temuan ini secara empiris mendukung asumsi teoretis bahwa informasi kontekstual yang lebih kaya pada respons yang lebih panjang meningkatkan keandalan penilaian kepribadian [8].

Namun, analisis per dimensi yang lebih mendalam mengungkapkan dinamika yang kompleks. Tabel 2 menyajikan perbandingan terfokus F1-Score dan nMCC untuk model berperforma terbaik pada setiap skenario panjang teks.

***Tabel 2**. F1-Score dan nMCC Per Dimensi untuk Model Terbaik*

Dimensi	Short Text (DistilBERT-Full)		Long Text (BERT-Full)	
	F1-Score	*nMCC*	*F1-Score*	*nMCC*
Openness	**0.8405**	0.5895	0.6952	**0.6436**
Conscientiousness	0.5872	**0.6954**	0.6403	0.6379
Extraversion	0.5641	0.6170	0.6602	0.6432
Agreeableness	0.6017	0.6642	**0.7198**	0.6162
Neuroticism	0.6424	0.6507	0.6549	0.5780

Catatan: F1-Score dan nMCC tertinggi untuk setiap kategori panjang teks dicetak tebal.

Perincian dimensional ini mengungkapkan dua pola kritis. Pertama, pada skenario teks pendek, terdapat perbedaan yang jelas antara kemudahan deteksi (detectability) dan keandalan (reliability). Dimensi Openness adalah yang paling mudah dideteksi dengan F1-Score luar biasa sebesar 0,8405, mengindikasikan bahwa penanda linguistiknya sangat eksplisit. Namun, dimensi yang paling andal diprediksi adalah Conscientiousness dengan skor nMCC tertinggi (0,6954), menunjukkan bahwa prediksi model untuk dimensi ini lebih seimbang dengan lebih sedikit kesalahan pada kasus positif maupun negatif.

Kedua, pada skenario teks panjang, konteks yang diperluas secara signifikan meningkatkan prediksi sifat yang berorientasi pada hubungan interpersonal. Agreeableness muncul sebagai dimensi dengan deteksi terbaik dengan F1-Score tertinggi (0,7198), mengindikasikan bahwa penanda linguistiknya bersifat halus dan memerlukan wacana yang lebih luas untuk diidentifikasi secara akurat. Sementara itu, Openness tetap menjadi dimensi yang paling andal diprediksi dengan nMCC tertinggi (0,6436), memperkuat bahwa sinyal linguistiknya secara konsisten kuat tanpa memandang panjang teks. Dampak diferensial panjang teks di berbagai dimensi kepribadian mengindikasikan bahwa ekspresi linguistik berbeda secara mendasar dari satu sifat ke sifat lainnya, sebuah pertimbangan kritis dalam mengembangkan sistem prediksi kepribadian yang andal [9].

4.3. Analisis Komparatif: BERT vs. DistilBERT dan Trade-off Efisiensi

Salah satu tujuan utama penelitian ini adalah mengevaluasi trade-off antara model yang besar dan mendasar (BERT) dengan versi yang lebih ringan dan terdistilasi (DistilBERT). Analisis menunjukkan bahwa DistilBERT tidak hanya menawarkan keunggulan efisiensi yang substansial, tetapi juga memberikan performa prediktif yang sangat kompetitif, dan dalam beberapa kasus lebih unggul.

Pada skenario teks pendek, DistilBERT dengan full fine-tuning menunjukkan keunggulan performa yang jelas atas BERT. Sebagaimana dirinci pada Tabel 1, DistilBERT mencapai skor tertinggi di seluruh metrik utama, termasuk Akurasi (0,6546 vs. 0,6313), F1-Score (0,6751 vs. 0,6600), dan terutama nMCC (0,6807 vs. 0,6615). Temuan ini signifikan karena mengindikasikan bahwa untuk tugas yang melibatkan input teks ringkas, arsitektur BERT yang lebih besar tidak serta merta menghasilkan performa yang lebih baik. Proses knowledge distillation tampaknya berhasil mempertahankan kemampuan semantik paling kritis dari BERT sekaligus menciptakan model yang lebih efisien dan efektif untuk konteks spesifik ini [6].

Pada skenario teks panjang, kesenjangan performa menyempit dan persaingan menjadi lebih bernuansa. Meskipun DistilBERT (Full) mencapai Akurasi keseluruhan yang sedikit lebih tinggi, BERT (Full) mengamankan skor nMCC tertinggi (0,6258 vs. 0,6231), mengindikasikan keandalan prediktif yang lebih unggul. Hal ini mengindikasikan bahwa parameterisasi BERT yang lebih besar memungkinkannya memanfaatkan sinyal kontekstual yang lebih kaya dan kompleks dalam teks panjang secara lebih efektif. Meskipun demikian, perbedaan performa ini bersifat marginal, terutama jika dipertimbangkan terhadap perbedaan biaya komputasi yang signifikan.

Keunggulan utama DistilBERT terletak pada efisiensi komputasinya. Data observasional yang dicatat selama eksperimen menunjukkan bahwa waktu pelatihan DistilBERT secara konsisten lebih singkat dibandingkan BERT pada pengaturan hyperparameter yang identik. Sebagai contoh, pada konfigurasi teks pendek yang optimal, DistilBERT menyelesaikan pelatihannya dalam sekitar 02:28:07, sedangkan BERT membutuhkan 04:05:26. Pengurangan durasi

pelatihan ini, dikombinasikan dengan jejak memori yang lebih kecil, menjadikan DistilBERT pilihan yang jauh lebih praktis untuk aplikasi dunia nyata di mana sumber daya komputasi dan latensi deployment merupakan kendala kritis. Hasil ini memposisikan DistilBERT sebagai pilihan optimal yang menawarkan keseimbangan luar biasa antara akurasi prediktif dan efisiensi operasional, sekaligus menantang asumsi bahwa model yang lebih besar selalu lebih baik. Hal senada juga ditemukan dalam penelitian berbasis Transformer lainnya, di mana model yang lebih ringan dengan fine-tuning yang tepat mampu mencapai performa yang kompetitif bahkan pada tugas klasifikasi teks yang kompleks [25].

4.4. Pembahasan

Hasil empiris penelitian ini menghasilkan beberapa wawasan kunci dalam tugas prediksi kepribadian secara komputasional. Temuan ini selaras dengan tinjauan sistematis terkini yang mengidentifikasi model Transformer pre-trained sebagai pendekatan yang dominan dan paling efektif dalam prediksi kepribadian berbasis teks modern [26]. Secara konsisten, penelitian ini menunjukkan keunggulan model deep learning atas metode rule-based tradisional, sekaligus menyoroti pentingnya panjang teks dalam meningkatkan akurasi prediksi serta adanya trade-off antara kompleksitas model dan efisiensi komputasi.

Pengamatan bahwa model berbasis Transformer secara signifikan mengungguli baseline berbasis leksikon selaras dengan tren dalam bidang pengenalan kepribadian komputasional [3]. Kemampuan model seperti BERT dan DistilBERT dalam memahami konteks linguistik terbukti menjadi faktor utama peningkatan performa, yang tidak dimiliki oleh pendekatan berbasis pencocokan kata kunci statis. Hal ini menguatkan bahwa untuk tugas dengan kebutuhan interpretasi bahasa yang kompleks, arsitektur deep learning tetap menjadi pendekatan state-of-the-art [12].

Penelitian ini juga memberikan analisis lebih mendalam mengenai trade-off antara strategi optimasi dan arsitektur model. Perbandingan antara head tuning dan full fine-tuning menunjukkan bahwa meskipun full fine-tuning menghasilkan performa terbaik, biaya komputasinya jauh lebih tinggi. Selain itu, stratifikasi berdasarkan panjang teks memberikan bukti empiris bahwa ketersediaan konteks linguistik secara langsung memengaruhi akurasi prediksi kepribadian [9].

Kontribusi praktis utama penelitian ini adalah demonstrasi efektivitas DistilBERT, yang mampu mencapai performa setara dengan BERT, bahkan lebih baik pada teks pendek, dengan waktu pelatihan yang lebih efisien. Namun demikian, penelitian ini memiliki keterbatasan, seperti penggunaan teks berbahasa Inggris dan formulasi klasifikasi biner untuk sifat kepribadian yang sebenarnya bersifat kontinu. Penelitian selanjutnya dapat mengembangkan pendekatan multibahasa, menggunakan metode regresi, serta mengevaluasi model bahasa yang lebih besar dan mutakhir.

5. KESIMPULAN

a. Penelitian ini menunjukkan bahwa model deep learning lebih unggul daripada pendekatan rule-based dalam prediksi kepribadian berbasis teks, dengan nMCC terbaik 0,6807 dibandingkan baseline 0,5483. Namun, tidak ada satu model yang paling dominan di semua kondisi, sehingga panjang teks terbukti memengaruhi performa. DistilBERT (Full Fine-Tuning) paling baik pada teks pendek, sedangkan BERT (Full Fine-Tuning) unggul pada teks panjang. Pada analisis dimensi, Openness paling mudah dideteksi, sementara Agreeableness lebih diuntungkan oleh konteks teks yang lebih panjang. Hasil ini juga menunjukkan bahwa Task-Specific Head Tuning dapat menjadi strategi yang efisien karena mampu menghasilkan performa yang kompetitif tanpa biaya komputasi sebesar Full Fine-Tuning.
b. Secara keseluruhan, DistilBERT menjadi pilihan paling seimbang antara akurasi dan efisiensi, tetapi pendekatan ini masih memiliki keterbatasan dalam konsistensi performa pada semua skenario. Penelitian selanjutnya dapat mengembangkan model yang lebih modern seperti LLM, memperkaya konteks dengan data multibahasa atau multimodal, serta memformulasikan masalah ini sebagai regresi agar prediksi kepribadian dapat ditangkap secara lebih kontinu dan bernuansa.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak terkait yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

[1] F. Celli dan B. Lepri, "Is Big Five better than MBTI?," dalam *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, Accademia University Press, 2019, hlm. 93–98. doi: 10.4000/books.aaccademia.3147.

[2] Y. Mehta, N. Majumder, A. Gelbukh, dan E. Cambria, "Recent Trends in Deep Learning Based Personality Detection," Agu 2019, doi: 10.1007/s10462-019-09770-z.

[3] G. Serapio-García *dkk.*, "Personality Traits in Large Language Models," Jun 2023, doi: 10.48550/arXiv.2307.00184.

[4] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, dan A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019. doi: 10.18653/v1/N19-1423.

[5] H. Christian, D. Suhartono, A. Chowanda, dan K. Z. Zamli, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *J. Big Data*, vol. 8, no. 1, Des 2021, doi: 10.1186/s40537-021-00459-1.

[6] V. Sanh, L. Debut, J. Chaumond, dan T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Mar 2020, doi: 10.48550/arXiv.1910.01108.

[7] N. Fatima, S. Gul, J. Ahmed, Z. H. Khand, dan G. Mujtaba, "A rule-based machine learning model for career selection through MBTI personality," *Mehran University Research Journal of Engineering and Technology*, vol. 41, no. 2, hlm. 185–196, Apr 2022, doi: 10.22581/muet1982.2202.18.

[8] N. Koenig *dkk.*, "Improving measurement and prediction in personnel selection through the application of machine learning," *Pers. Psychol.*, vol. 76, no. 4, hlm. 1061–1123, Des 2023, doi: 10.1111/peps.12608.

[9] E. Kerz, Y. Qiao, S. Zanwar, dan D. Wiechmann, "Pushing on Personality Detection from Verbal Behavior: A Transformer Meets Text Contours of Psycholinguistic Features," 2022. doi: 10.18653/v1/2022.wassa-1.17.

[10] Dr. V Geetha, D. C. K. Gomathy, Mr. D. S. D. Vallab Yaratha Yagn, dan S. Praneesh, "THE ROLE OF NATURAL LANGUAGE PROCESSING," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 07, no. 11, hlm. 1–11, Nov 2023, doi: 10.55041/IJSREM27094.

[11] A. Bittermann dan A. Fischer, "Natural Language Processing in Psychology," 1 Juli 2024, *Hogrefe Verlag GmbH & Co. KG*. doi: 10.1027/2151-2604/a000568.

[12] Y. Lecun, Y. Bengio, dan G. Hinton, "Deep learning," 27 Mei 2015, *Nature Publishing Group*. doi: 10.1038/nature14539.

[13] R. R. Mccrae dan O. P. John, "An Introduction to the Five-Factor Model and Its Applications," 1992. doi: 10.1111/j.1467-6494.1992.tb00970.x.

[14] M. Sokolova dan G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, hlm. 427–437, Jul 2009, doi: 10.1016/j.ipm.2009.03.002.

[15] D. Chicco dan G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, Jan 2020, doi: 10.1186/s12864-019-6413-7.

[16] M. Gjurković, M. Karan, I. Vukojević, M. Bošnjak, dan J. Šnajder, "PANDORA Talks: Personality and Demographics on Reddit," Jun 2021, doi: 10.18653/v1/2021.socialnlp-1.12.

[17] F. Celli, F. Pianesi, D. Stillwell, dan M. Kosinski, "Workshop on Computational Personality Recognition: Shared Task," 2021. doi: 10.1609/icwsm.v7i2.14467.

[18] J. W. Pennebaker dan L. A. King, "Linguistic Styles: Language Use as an Individual Difference," 1999. doi: 10.1037/0022-3514.77.6.1296.

[19] H. Pouransari *dkk.*, "Dataset Decomposition: Faster LLM Training with Variable Sequence Length Curriculum," Jan 2025, doi: 10.48550/arXiv.2405.13226.

[20] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, dan F. L. Wang, "Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment," Des 2023, doi: 10.48550/arXiv.2312.12148.

[21] Z. Yang, M. Ding, Y. Guo, Q. Lv, dan J. Tang, "Parameter-Efficient Tuning Makes a Good Classification Head," Mar 2023, doi: 10.48550/arXiv.2210.16771.

[22] J. M. Digman, "PERSONALITY STRUCTURE: EMERGENCE OF THE FIVE-FACTOR MODEL," 1990. doi: 10.1146/annurev.ps.41.020190.002221.

[23] S. C. Ma, T. Ermakova, dan B. Fabian, "FairGridSearch: A Framework to Compare Fairness-Enhancing Models," dalam *Proceedings - 2023 22nd IEEE/WIC*

International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2023, Institute of Electrical and Electronics Engineers Inc., 2023, hlm. 388–394. doi: 10.1109/WI-IAT59888.2023.00064.

[24] A. Chowanda, D. Suhartono, E. W. Andangsari, dan K. Z. bin Zamli, "MACHINE LEARNING ALGORITHMS EXPLORATION FOR PREDICTING PERSONALITY FROM TEXT," *ICIC Express Letters*, vol. 16, no. 2, hlm. 117–125, Feb 2022, doi: 10.24507/icicel.16.02.117.

[25] Divaretta Kiesa Sumartha, "IMPLEMENTASI INDOBERT UNTUK ANALISIS SENTIMEN OPINI PUBLIK TERHADAP KEBIJAKAN KENAIKAN UKT DI ERA PEMERINTAHAN," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 3S1, Okt 2025, doi: 10.23960/jitet.v13i3S1.7880.

[26] K. Kelvin dan Y. Utomo, "Overview of Text Based Personality Prediction Using Deep Learning," *Engineering, MAthematics and Computer Science Journal (EMACS)*, vol. 6, no. 2, hlm. 93–100, Mei 2024, doi: 10.21512/emacsjournal.v6i2.11550.