e-ISSN: 2579-3772
doi: 10.61769/telematika.v20i2.775

Analisis Pengaruh Karakteristik Masukan Teks terhadap Kinerja MiniLM v2-L6-H384 dan BERT-*Base-Uncased* pada *Quora Question Pairs*

Ken Ratri Retno Wardani[1], Inge Martina[2], Jimmy Fong Xin Wern[3]

[1-3]*Program Studi Informatika, Institut Teknologi Harapan Bangsa,*
Jl. Dipati Ukur No.80-84 Bandung, Jawa Barat, Indonesia
[2]inge@ithb.ac.id
[3]jimmyfsec1@gmail.com

*Korespondensi: [1]ken_ratri@ithb.ac.ic

Abstract— Knowledge distillation is a technique to simplify large language models into smaller models while maintaining accuracy. The Bidirectional Encoder Representations from Transformers (BERT) offers strong performance but requires significant computational resources, whereas the Mini Language Model (MiniLM) is five times smaller. This study aims to compare the performance of these models on the Quora Question Pairs dataset, focusing on the influence of sequence length and token rarity on classification accuracy. Both models were fine-tuned using identical training parameters. Results show that BERT achieved 91.22% accuracy and an F1-score of 88.17%, slightly outperforming MiniLM which reached 90.12% accuracy and an F1-score of 86.73%. However, MiniLM delivered 5.3 times faster inference. Variable analysis indicates that accuracy improves with longer sequences and rarer tokens, where BERT is more effective at capturing subtle semantic nuances. These findings provide empirical guidance for model optimization in real-time systems, where MiniLM's efficiency is acceptable with minimal accuracy loss. Future research is encouraged to explore hybrid approaches that utilize both small and large models based on input complexity.

Keywords— knowledge distillation, MiniLM, BERT, semantic equivalence, quora question pairs, sequence length, token rarity.

Abstrak— Knowledge distillation* merupakan teknik untuk menyederhanakan model bahasa besar menjadi model yang lebih ringkas dengan tetap mempertahankan akurasi. Bidirectional encoder representations from transformers (BERT) menawarkan kinerja kuat, namun memerlukan sumber daya komputasi besar, sedangkan mini language model (MiniLM) memiliki ukuran model lima kali lebih kecil. Penelitian ini bertujuan membandingkan kinerja kedua model tersebut pada dataset quora question pairs dengan fokus pada pengaruh *sequence length* dan kelangkaan token (*token rarity*) terhadap akurasi klasifikasi. Kedua model dilatih menggunakan parameter pelatihan yang identik. Hasil pengujian menunjukkan BERT mencapai akurasi 91,22% dan F1-score 88,17%, sedikit lebih unggul dari MiniLM yang mencapai akurasi 90,12% dan F1-score 86,73%. Namun, MiniLM memberikan kecepatan inferensi 5,3 kali lebih cepat. Temuan ini memberikan panduan empiris untuk optimalisasi model untuk lingkungan dengan keterbatasan sumber daya komputasi atau kebutuhan respons *real-time*, di mana efisiensi MiniLM dapat diterima dengan sedikit penurunan akurasi. Penelitian mendatang disarankan untuk mengeksplorasi sistem hibrida yang mengalihkan tugas kompleks ke model besar dan tugas umum ke model yang lebih kecil.

***Kata Kunci*— distilasi pengetahuan, MiniLM, BERT, kesamaan semantik, *quora question pairs*, panjang sekuens, kelangkaan token.**

I. PENDAHULUAN

Perkembangan model bahasa berbasis transformer [1] seperti *bidirectional encoder representations from transformers* (BERT) [2], [3] telah merevolusi deteksi kesamaan semantik, khususnya pada tugas *paraphrase identification*, seperti *quora question pairs* (QQP). Namun, model-model besar ini menghadapi kendala signifikan dalam implementasi praktis, termasuk kebutuhan sumber daya komputasi yang tinggi dan latensi inferensi yang besar, sehingga kurang sesuai untuk lingkungan dengan keterbatasan sumber daya atau aplikasi yang menuntut respons waktu nyata

Knowledge distillation (KD) muncul sebagai pendekatan efektif untuk mentransfer kapabilitas model "*teacher*" berukuran besar ke model "*student*" yang lebih ringkas [4], seperti MiniLM [5], [6]. Relevansi pendekatan ini tetap bertahan hingga era model bahasa besar modern, sebagaimana ditunjukkan oleh kerangka kerja distilasi terkini yang berfokus pada efisiensi dan transfer kemampuan penalaran, seperti Orca [7] yang mendistilasi penalaran kompleks, serta TinyLlama [8] yang menjadi model student utama dalam menjembatani efisiensi dengan kecerdasan setingkat model *frontier*. Meskipun demikian, sebagian besar penelitian sebelumnya mengevaluasi model hasil distilasi berdasarkan metrik kinerja agregat sehingga masih terdapat kesenjangan empiris terkait bagaimana karakteristik masukan teks secara spesifik memengaruhi akurasi model distilasi dibandingkan model aslinya. Sebagian besar ulasan literatur yang ada saat ini lebih berfokus pada skor agregat tanpa membedah variabel internal teks [9]-[12].

Penelitian ini mengisi celah tersebut melalui analisis kuantitatif terhadap performa model pada dataset QQP. Kontribusi utama penelitian ini meliputi:

1. perbandingan sistematis antara BERT-*base-uncased* sebagai "*teacher*" dan MiniLMv2-L6-H384 sebagai "*student*" pada dataset QQP berdasarkan akurasi, F1-*score*, dan kecepatan inferensi, serta
2. analisis empiris pengaruh panjang urutan teks (*sequence length*) dan kelangkaan token (token *rarity*)—yang didefinisikan berdasarkan distribusi token korpus—terhadap keputusan klasifikasi model.

Tujuan dari penelitian ini adalah memetakan *trade-off* antara ukuran model dan performa klasifikasi, serta mengidentifikasi batasan model kecil dalam menangani kompleksitas semantik teks dibandingkan model transformer.

II. METODOLOGI

Bab ini menjelaskan metodologi komparatif yang menggabungkan metrik kuantitatif dengan analisis perilaku model secara kualitatif. Prosedur ini berfokus pada evaluasi dampak knowledge distillation terhadap performa klasifikasi semantik pada arsitektur BERT dan MiniLM.

A. Desain dan Prosedur Penelitian

Penelitian ini dirancang sebagai studi komparatif dengan pendekatan kualitatif berbasis analisis perilaku model yang bertujuan untuk mengevaluasi dampak *knowledge distillation* terhadap performa model bahasa berbasis transformer. Desain penelitian ini berfokus pada perbandingan karakteristik kinerja antara model BERT sebagai *teacher model* dan MiniLM sebagai *student model* pada tugas *semantic equivalence classification*. Evaluasi dilakukan menggunakan dataset *quora question pairs* (QQP) [12] dengan menelaah bagaimana perbedaan arsitektur dan ukuran model memengaruhi keputusan klasifikasi dalam berbagai kondisi linguistik. Pendekatan ini memungkinkan analisis mendalam terhadap pola keberhasilan dan kegagalan model dalam menangkap kesamaan makna yang tidak sepenuhnya tercermin melalui metrik evaluasi agregat semata.

Variabel penelitian didefinisikan sebagai berikut:

1. Variabel independen adalah jenis arsitektur model transformer, yaitu BERT sebagai *teacher model* dan MiniLM sebagai *student model*, perbedaan arsitektur ini secara inheren mencakup perbedaan jumlah parameter dan metode pelatihan, yakni *fine-tuning* langsung pada BERT dan *knowledge distillation* pada MiniLM.
2. Variabel dependen adalah metrik evaluasi kinerja model yang diukur menggunakan akurasi dan F1-*score* pada tugas *semantic equivalence classification*.
3. Variabel kontrol adalah *hyperparameter* pelatihan (*learning rate*, *batch size*, *optimizer*, *jumlah epochs*), skema tokenisasi, serta rasio pembagian data latih, validasi, dan uji.

Prosedur penelitian dilaksanakan secara kronologis sebagai berikut:

1) *Pengumpulan dan Pemilihan Data*: Dataset QQP (D) diunduh dari repositori resmi HuggingFace. Dataset ini berisi pasangan pertanyaan (q_1, q_2) berlabel $y \in \{0, 1\}$ yang menunjukkan kesetaraan semantik.

2) *Pembagian Dataset:* Data dibagi menjadi subset pelatihan (D_{train}), validasi (D_{val}), dan pengujian (D_{test}) dengan rasio 45,7% : 5,1% : 49,2%, memastikan distribusi label tetap seimbang.

3) *Preprocessing:* Setiap pasangan pertanyaan ditokenisasi menggunakan BERT *tokenizer* berbasis *WordPiece vocabulary*. *Sequence length* dibatasi hingga 512 *token* dengan penerapan *padding* dan *attention mask* yang sesuai.

4) Konfigurasi Model:

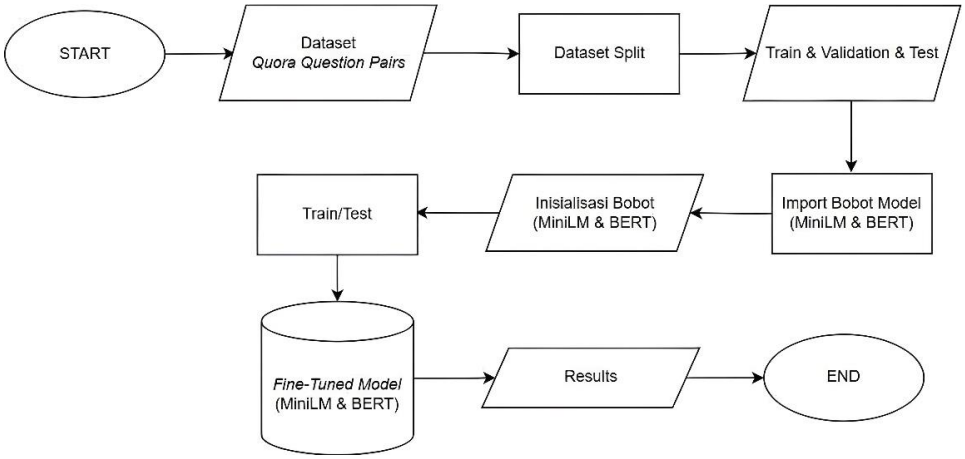
a. Model *teacher* menggunakan BERT, $\approx$ 109M parameter, 12 *encoder layers*, *hidden size* 768.
b. Model *student* menggunakan MiniLM, $\approx$ 22M parameter, 6 *encoder layers*, *hidden size* 384, hasil *attention-based knowledge distillation* dari *teacher*.

5) *Fine-tuning:* Kedua model dilatih pada D_{train} menggunakan konfigurasi identik, yaitu AdamW *optimizer*, *learning rate* 2×10^{-5}, *batch size* 32, dan 5 *epochs*.

6) *Evaluasi:* Evaluasi kinerja dilakukan pada D_{val} setelah setiap *epoch* untuk memantau stabilitas pelatihan serta pada D_{test} untuk memperoleh hasil akhir dengan menggunakan metrik akurasi dan F1-*score*.

7) *Analisis Komparatif Kualitatif:* Hasil prediksi model *teacher* (T) dan *student* (S) dibandingkan secara kualitatif untuk mengidentifikasi pola kesalahan yang berkaitan dengan variasi panjang urutan (*sequence length*) dan kelangkaan token (*low-frequency tokens*).

Gambar 1 menunjukkan urutan proses global dari penelitian ini.



Gambar 1 Urutan proses global

B. Semantic Equivalence Analysis

Semantic equivalence analysis adalah teknik analisis teks yang bertujuan untuk menentukan apakah dua unit bahasa—seperti kalimat, pertanyaan, atau dokumen pendek—menyampaikan makna yang setara secara semantik, meskipun terdapat perbedaan dalam struktur sintaksis, pilihan kosakata, atau gaya bahasa yang digunakan [12]. Pendekatan ini tidak hanya bergantung pada kesamaan leksikal, tetapi juga mempertimbangkan hubungan kontekstual dan representasi makna laten yang terkandung dalam teks. Dalam praktiknya, hasil analisis ini pada tingkat dataset umumnya direpresentasikan dalam bentuk label biner (setara atau tidak setara), sementara pada tingkat keluaran model sering dinyatakan sebagai nilai probabilitas yang menunjukkan tingkat keyakinan model terhadap kesetaraan makna antar kalimat. Analisis *semantic equivalence* memiliki peran penting dalam deteksi duplikasi konten dan sistem tanya-jawab karena dapat membantu memastikan bahwa informasi yang berbeda penyajiannya tetap diakui sebagai memiliki makna yang sama untuk mengurangi instansi redundan.

C. Text Preprocessing

Dalam arsitektur transformer, *tokenization* memiliki peran yang sangat penting sebagai satu-satunya tahap *preprocessing* yang esensial. Hal ini disebabkan oleh kemampuan mekanisme *self-attention* dalam transformer yang memungkinkan model untuk mempelajari hubungan kontekstual antar token secara langsung, termasuk antar token yang berjauhan dalam suatu kalimat. Dengan kemampuan tersebut, transformer tidak lagi memerlukan tahap *preprocessing* linguistik yang kompleks, seperti *stemming* atau *lemmatization*, yang umum digunakan pada model berbasis aturan atau pembelajaran mesin tradisional. Sebagai gantinya, representasi *subword* yang dihasilkan melalui proses *tokenization* sudah memadai untuk menangkap variasi morfologis dan makna kontekstual dalam teks.

D. BERT dan MiniLM

BERT adalah model berbasis arsitektur *encoder-only* transformer (lihat Gambar 2). Arsitektur ini memproses urutan masukan $X = (x_1, x_2, \ldots, x_n)$ secara dua arah melalui mekanisme *multi-head self-attention* yang menghasilkan representasi kontekstual $\mathbf{H} \in \mathrm{R}^{n \times d_{\text{model}}}$. Di sini, n menyatakan panjang urutan (jumlah token dalam masukan), d_{model} adalah dimensi vektor representasi tiap token, dan $\mathbf{H}$ merupakan matriks yang menyimpan representasi kontekstual untuk seluruh token dalam urutan. Setiap token x_i memiliki akses penuh terhadap informasi kontekstual dari token lain. Pemrosesan ini berbeda dengan model sekuensial tradisional (misalnya LSTM dan RNN standar) yang bersifat unidireksional karena transformer menghilangkan keterbatasan ketergantungan arah dan mampu menangkap relasi semantik jarak jauh secara efisien [1], [2].

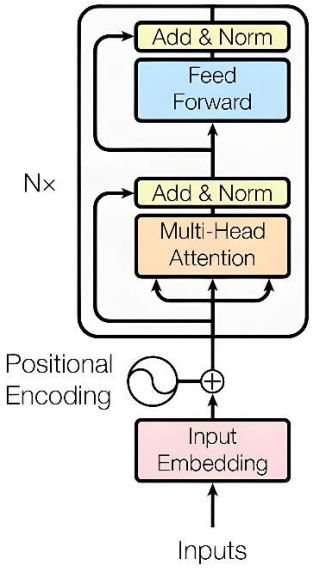
MiniLM merupakan versi kompak dari BERT yang diperoleh melalui proses *knowledge distillation* (KD). Dalam paradigma KD terdapat dua entitas model: *teacher* f^T dengan kapasitas besar, dan *student* f^S yang lebih ringan. Model *student* dilatih untuk mereplikasi fungsi pemetaan semantik yang dihasilkan oleh *teacher* pada himpunan data D, dengan *loss function*:

$$\mathcal{L}_{\mathrm{K}} = \sum_{e \in \mathcal{D}} L\big(f^S(e), f^T(e)\big) \tag{1}$$

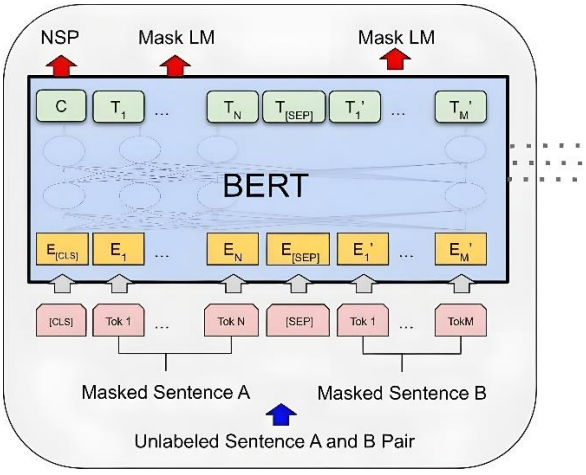
e merepresentasikan sebuah entitas data (misalnya pasangan teks), $\mathrm{L}(\cdot,\cdot)$ adalah *loss function* (misalnya *cross entropy* atau *KL-divergence*) yang mengukur jarak antara keluaran *student* dan *teacher* Formulasi ini bersifat *model agnostic* sehingga dapat digunakan untuk distilasi berbasis *output logits*, distribusi probabilitas atau berbagai representasi internal [4].

Tahap fundamental dalam pengembangan BERT dan MiniLM adalah *pre-training* [2], [5]. Model dilatih pada korpus teks yang besar untuk memahami struktur dan semantik bahasa. BERT memperkenalkan dua tujuan *pre-training* utama, yaitu: *masked language modeling* (MLM) dan *next sentence prediction* (NSP). MLM melatih model untuk memprediksi kata yang disembunyikan berdasarkan konteks sekitarnya, sedangkan NSP melatih model untuk memprediksi apakah dua kalimat saling melanjutkan, lihat Gambar 3. MiniLM, di sisi lain, menggunakan teknik distilasi pengetahuan untuk mentransfer kemampuan dari model *teacher* yang lebih besar ke model *student* yang lebih kecil. Ini bisa dilakukan dengan meniru prediksi token atau representasi internal yang dihasilkan oleh model *teacher* pada tugas *pre-training*, seperti MLM.

Untuk menerapkan model yang sudah dilatih ini pada tugas spesifik, seperti klasifikasi teks atau menjawab pertanyaan, dilakukan proses *fine-tuning*. Dalam proses ini, lapisan *output* khusus ditambahkan pada bagian atas model yang sudah dilatih. Seluruh model, termasuk lapisan *output* baru, kemudian dilatih kembali menggunakan data berlabel yang spesifik untuk tugas tersebut. *Fine-tuning* memungkinkan model



Gambar 2 Arsitektur *encoder-only* transformer [2]



Gambar 3 Tujuan *pre-training*: MLM dan NSP [2]

untuk mengadaptasi pengetahuan umum yang sudah didapatkan saat *pre-training* ke dalam konteks tugas yang lebih spesifik. Proses ini membutuhkan daya komputasi yang jauh lebih kecil dan dapat dilakukan dengan dataset yang lebih sedikit dibandingkan saat *pre-training*. Fleksibilitas dan efisiensi *fine-tuning* ini menjadikan BERT dan MiniLM sangat efektif untuk berbagai tugas pemrosesan bahasa alami.

E. Pembahasan Dataset

Dataset QQP digunakan dalam penelitian ini untuk tugas klasifikasi pasangan pertanyaan, di mana model harus menentukan apakah dua pertanyaan memiliki makna yang setara (dilabeli '*duplicate*') atau tidak ('*non-duplicate')*. Dataset ini terdiri dari total 795.241 pasangan pertanyaan yang terbagi ke dalam tiga subset, *train*, *validation*, dan *test*. *Subset train* berisi 363.846 pasangan pertanyaan dengan 229.468 pasangan berlabel *non-duplicate* dan 134.378 berlabel *duplicate*. *Subset validation* berisi 40.430 pasangan pertanyaan, sedangkan *subset* test terdiri dari 390.965 pasangan pertanyaan tanpa label yang seharusnya digunakan untuk evaluasi akhir. Dalam kerangka GLUE *benchmark* [12] yang mewajibkan pengiriman bobot model serta hasil dari seluruh pengujian. Mengingat penelitian ini difokuskan secara spesifik pada tugas QQP tanpa mengikuti keseluruhan protokol GLUE, maka evaluasi kinerja model dalam penelitian ini dilakukan menggunakan subset *validation* yang tidak terlibat dalam proses pelatihan. Struktur data pada dataset ini terdiri dari empat field utama: *question1* dan *question2* yang berisi pasangan pertanyaan, label yang merupakan indikator biner (1 untuk duplikat, 0 untuk non-duplikat), dan idx sebagai indeks unik.

1) Distribusi dalam Dataset: Analisis distribusi frekuensi token pada dataset QQP menunjukkan adanya ketidakseimbangan yang signifikan. Dataset ini mencakup sebanyak 25.677 token unik, di mana sebagian besar token muncul dengan frekuensi yang sangat rendah, sementara sejumlah kecil token umum—seperti "*best*" dan "*get*"—muncul dengan frekuensi yang sangat tinggi. Skor token *rarity* dihitung menggunakan -log(frekuensi token relatif), lihat persamaan (2), di mana skor yang lebih tinggi menunjukkan token yang lebih langka. Ketidakseimbangan ini berpotensi mempengaruhi kinerja model, khususnya pada pertanyaan yang mengandung token langka, karena token tersebut memiliki representasi yang lebih terbatas dalam data pelatihan.

2) *Token Rarity:* Kelangkaan token dalam dataset diukur menggunakan metrik *token rarity score*, yang dirumuskan sebagai:

$$\text{rarity}(t) = -\log\left(\frac{\text{count}(t)}{\sum_{t'} \text{count}(t')}\right) \tag{2}$$

di mana *count*(*t*) adalah jumlah kemunculan token *t* dalam seluruh korpus, dan penyebut merupakan total kemunculan seluruh token. Skor yang lebih tinggi menunjukkan bahwa token tersebut jarang muncul dan kemungkinan memiliki representasi distribusional yang berpotensi kurang terasah pada embedding space hasil pelatihan.

3) *Catatan Dataset:* Dataset QQP memiliki tingkat kompleksitas yang berasal akibat ambiguitas bahasa alami. Satu kalimat dapat mengandung berbagai kemungkinan makna bergantung pada konteks penggunaannya. Kondisi ini menyebabkan hubungan semantik antarkalimat bersifat *many-to-many*, yaitu dua kalimat yang berbeda dapat merujuk pada makna yang sama maupun berbeda secara bersamaan (Kalimat 1 -> Makna ≠ Makna <- Kalimat 2). Namun demikian, proses anotasi dalam dataset QQP mereduksi kompleksitas semantik tersebut ke dalam skema pelabelan biner (0 untuk *non-duplicate* dan 1 untuk *duplicate*) yang hanya merepresentasikan satu keputusan diskret atas spektrum makna yang sejatinya bersifat kontinu. Meskipun pendekatan ini praktis untuk keperluan pemodelan dan evaluasi, reduksi tersebut berpotensi menyederhanakan nuansa semantik yang ada. Selain itu, variasi penilaian antar anotator

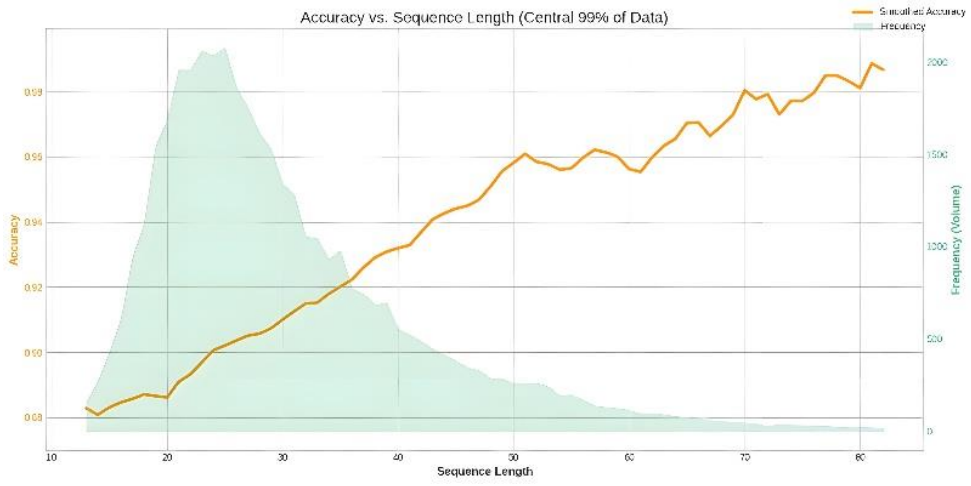
turut menjadi sumber ketidakpastian, karena tidak adanya ambang batas numerik yang eksplisit dan konsisten dalam menentukan tingkat kesetaraan makna. Akibatnya, pasangan pertanyaan yang sama dapat menerima label yang berbeda, bergantung pada ambang ekuivalensi semantik yang diterapkan oleh masing-masing annotator.

III. HASIL DAN PEMBAHASAN

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa peningkatan *sequence length* berkorelasi positif dengan akurasi model karena panjang urutan yang lebih besar menyediakan konteks semantik yang lebih kaya untuk proses pengambilan keputusan. Selain itu, berlawanan dengan anggapan umum bahwa token langka cenderung menghambat kinerja model, temuan empiris dalam penelitian ini menunjukkan bahwa *token rarity* justru dapat berfungsi sebagai penanda kontekstual yang informatif. Keberadaan token langka membantu model membedakan makna secara lebih spesifik sehingga berkontribusi pada peningkatan akurasi, baik pada model BERT maupun MiniLM yang diuji menggunakan dataset QQP.

A. Hasil Pengujian Maximum Sequence Length

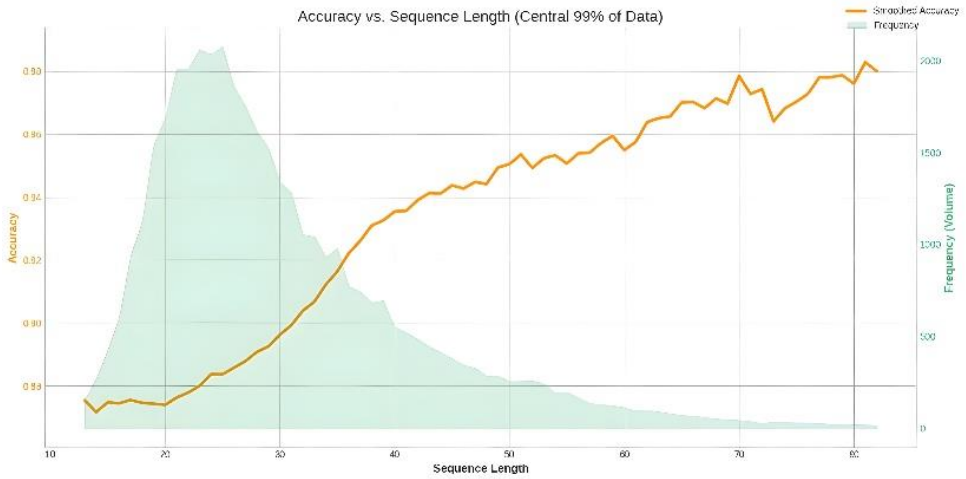
Peningkatan akurasi pada kedua model, BERT dan MiniLM, menunjukkan korelasi positif dengan bertambahnya *sequence length* yang menyediakan konteks semantik yang lebih kaya dan mendukung pemahaman semantik. Pada rentang *sequence length* 20–60, BERT menunjukkan kenaikan akurasi yang lebih cepat dan cenderung linear (lihat Gambar 4 dan 5), yang dapat dikaitkan dengan kedalaman arsitekturnya yang lebih besar sehingga mampu memanfaatkan tambahan konteks secara lebih efektif. Sebaliknya, MiniLM memerlukan *sequence length* yang lebih panjang untuk mencapai tingkat akurasi yang setara, yang mencerminkan keterbatasan kapasitas representasi akibat ukuran model yang lebih kecil. Pada *sequence length* di atas 90, kedua model mendekati akurasi sempurna, namun kondisi ini kemungkinan dipengaruhi oleh keterbatasan jumlah data pada rentang tersebut sehingga meningkatkan risiko *overfitting*, mengingat mayoritas data berada pada rentang *sequence length* 20–40. Namun, perlu dicatat bahwa Baillargeon dan Lamontagne [13] menemukan bahwa model transformer dapat mengeksploitasi panjang sekuens sebagai jalan pintas dalam klasifikasi; pada penelitian ini, hasil peningkatan akurasi diasumsikan mencerminkan pemahaman konteks yang lebih mendalam.



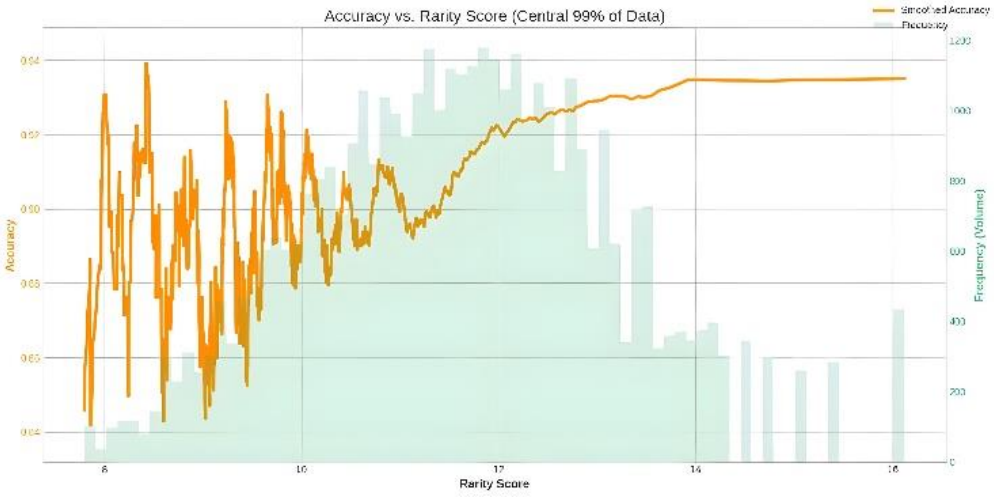
Gambar 4 Hubungan antara *sequence length* (sumbu-X) dan akurasi model (sumbu-Y kiri, oranye) serta frekuensi kemunculan sekuens (sumbu-Y kanan, hijau) pada data untuk BERT

B. Hasil Pengujian Token Rarity

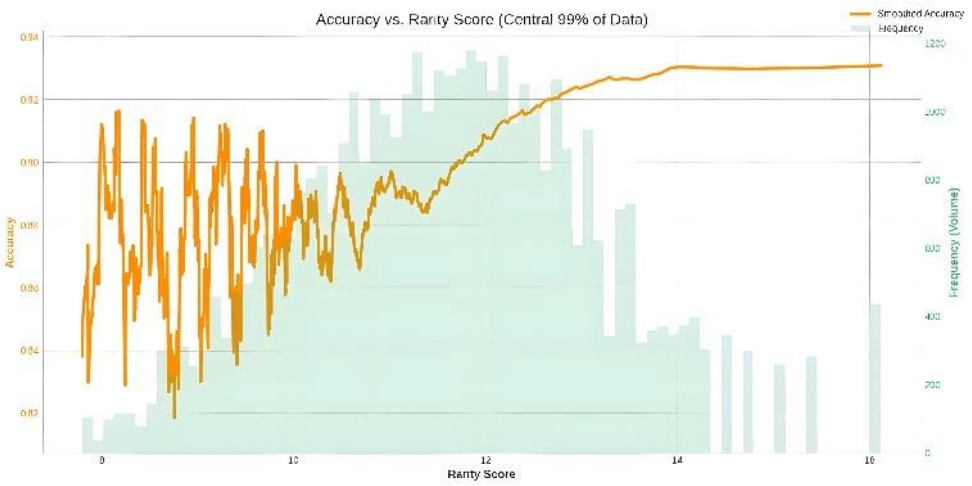
Grafik hasil eksperimen menunjukkan bahwa akurasi BERT dan MiniLM berfluktuasi dan cenderung lebih rendah pada token yang sangat umum yang ditandai dengan skor *token rarity* pada rentang 8–10. Kondisi ini kemungkinan disebabkan oleh tingkat ambiguitas yang lebih tinggi pada token-token umum tersebut. Sebaliknya, akurasi kedua model meningkat dan menjadi lebih stabil, mendekati nilai 0,94, pada token yang lebih langka dengan skor token *rarity* di atas 11. Temuan ini mengindikasikan bahwa token yang jarang muncul membawa informasi kontekstual yang lebih spesifik sehingga membantu model dalam membedakan makna dan menghasilkan prediksi yang lebih akurat. Meskipun BERT lebih stabil pada token frekuensi menengah, kedua model mencapai akurasi maksimum yang serupa pada token yang informatif. Ini memperkuat temuan bahwa *token rarity* signifikan mempengaruhi performa model (lihat Gambar 6 dan 7). Berbeda dengan temuan Yu et al. [14] melalui Dict-BERT yang menyoroti token langka sebagai kelemahan karena representasinya kurang terlatih, hasil pada dataset QQP justru menunjukkan bahwa token langka dapat berperan sebagai penanda kontekstual yang membantu meningkatkan akurasi model.



Gambar 5. Visualisasi seperti Gambar 4, namun untuk model MiniLM



Gambar 6. Hubungan antara *token rarity* (sumbu-X) dan akurasi model (sumbu-Y kiri, oranye) serta frekuensi kemunculan sekuens (sumbu-Y kanan, hijau) pada data untuk BERT



Gambar 7. Visualisasi seperti Gambar 6, namun untuk model MiniLM

C. Pembahasan Hasil

Berdasarkan pengujian, BERT secara umum memiliki akurasi lebih tinggi (91,22%) dibandingkan MiniLM (90,12%) yang sejalan dengan perbedaan jumlah parameter yang signifikan. F1-*score* secara keseluruhan juga menunjukkan pola serupa dengan BERT memperoleh 88,17% dan MiniLM 86,73%. Dalam hal *sequence length*, kedua model menunjukkan bahwa sekuens yang sangat pendek menghilangkan informasi dan menyebabkan banyaknya *padding*. Meskipun demikian, *attention mask* membantu kedua model tetap efisien, meskipun BERT terbukti lebih tangguh terhadap penambahan *sequence length*. Terkait *token rarity*, BERT menunjukkan kinerja yang lebih stabil pada token langka berkat *pre-training* yang luas. Sementara MiniLM, meskipun memiliki akurasi fluktuatif pada token umum, dapat menyamai BERT pada token yang sangat spesifik. Hal ini menunjukkan bahwa MiniLM adalah alternatif yang efisien untuk kasus-kasus umum dengan keterbatasan sumber daya. Kesimpulannya, ada *trade-off* antara ukuran model, efisiensi, dan akurasi. BERT lebih stabil, sedangkan MiniLM unggul dalam efisiensi dengan performa yang sangat baik untuk data dengan distribusi umum.

D. Analisis Kesalahan Kualitatif

Analisis kesalahan ini dilakukan terhadap 50 pasangan pertanyaan yang merupakan sampel kesalahan yang dihasilkan oleh model MiniLM dan/atau BERT. Pendekatan yang digunakan mengelompokkan kesalahan ke dalam pola tematik berdasarkan karakteristik yang muncul. Persentase yang dilaporkan merefleksikan proporsi kemunculan pola dalam kumpulan sampel kesalahan.

1) *Pengaruh Panjang Kalimat:* Sekitar 30% dari kasus kesalahan berkaitan dengan panjang kalimat. Kalimat panjang mempersulit deteksi parafrasa jika struktur sintaksis berubah drastis. Menariknya, *sequence length* justru membantu model membedakan makna lebih baik, memungkinkan penanganan kalimat panjang secara akurat saat perbedaannya sangat jelas. Contoh:

1. *"How's hostel life?" vs. "How is your hostel life?"* Label: 0 (Tidak Duplikat). Kedua model salah memprediksi (1). Perbedaan makna pada kata "*your*" sulit ditangkap tanpa konteks.
2. *"I am 25 year old guy and never had a girlfriend. Is this weird?" vs. "I am 25 years old. I have never had a girlfriend. Is something wrong with me?"* Label: 1 (Duplikat). Kedua model (MiniLM dan BERT) memprediksi dengan benar (1). Kedua pertanyaan menyampaikan ketidakpastian atas pengalaman yang sama (tidak pernah memiliki pacar di usia 25) dan, meskipun menggunakan kata-kata berbeda, model berhasil menangkap kesamaan makna.

2) *Token Rarity dan Entitas Tidak Umum:* Sekitar 10% dari kesalahan melibatkan token frekuensi rendah atau entitas tidak umum. Token frekuensi rendah atau entitas langka sering dianggap tantangan. Namun, token langka ini justru membantu model membedakan pertanyaan. Kedua model menunjukkan ketahanan, khususnya saat konteks utama tidak bergantung pada token langka tersebut. Contoh: *"What is the Sahara, and how do the average temperatures there compare to the ones in the Patagonian Desert"? vs. "...in the Registan Desert?"* Label: 1 (Duplikat). Kedua model benar (1). Ini menunjukkan bahwa model tangguh terhadap perubahan nama gurun yang spesifik.

3) *Perubahan Nuansa:* Pola ini muncul pada sekitar 20% dari kesalahan, model cenderung gagal ketika dua pertanyaan memiliki banyak kata sama namun makna atau intensi berbeda. BERT lebih baik dalam membedakan nuansa ini. Contoh: *"Why is my life getting so complicated?" vs. "Why is my life so complicated?"* Label: 0 (Tidak Duplikat). BERT benar (0), membedakan antara proses (”getting so complicated”) dan kondisi saat ini (”so complicated”). MiniLM salah (1).

4) *Sinonim dan Reformulasi Kalimat Pendek:* Kasus ini mencakup sekitar 15% dari keseluruhan data yang dianalisis. Dalam kalimat pendek, model umumnya berhasil menangkap sinonim atau parafrasa yang membuat pertanyaan duplikat, terutama jika sinonimnya umum. Contoh: *"What was the deadliest battle in history?" vs."What was the bloodiest battle in history?"* Label: 1 (Duplikat). Kedua model benar (1).

5) *Pengetahuan Spesifik dan Perubahan Nama Entitas:* Pola ini muncul pada sekitar 10% dari sampel kesalahan. Model, terutama yang lebih kecil, dapat mengabaikan modifikasi kecil. Namun, pada contoh ini, kedua model tepat. Contoh: *"Why are African-Americans so beautiful?" vs. "Why are hispanics so beautiful?"* Label: 0 (Tidak Duplikat). Kedua model benar (0).

6) *Kasus Ambigu dan Tantangan Demarkasi:* Sekitar 15% dari kasus kesalahan menunjukkan kasus di mana kedua model gagal. Ini mengindikasikan label yang ambigu. Contoh: *"What are the best things to do in Hong Kong?" vs. "What is the best thing in Hong Kong?"* Label: 1 (Duplikat). Kedua model salah (0). Perbedaan semantik antara ”*best things to do*” dan ”*best thing*” mungkin dianggap signifikan, walaupun intensi pengguna mungkin bisa serupa.

Kasus ini menyoroti permasalahan ambiguitas dalam pelabelan dataset, khususnya dalam membedakan pasangan pertanyaan yang benar-benar duplikat dengan yang tidak. Pasangan pertanyaan tersebut secara semantik menunjukkan tingkat kemiripan yang tinggi, sehingga lebih tepat diklasifikasikan sebagai *probable duplicate*.

IV. SIMPULAN

Penelitian ini menunjukkan bahwa MiniLM mampu mencapai performa yang kompetitif pada tugas *semantic equivalence classification*, sekaligus menawarkan keuntungan signifikan dalam kecepatan inferensi dibandingkan BERT. Temuan analisis terhadap *sequence length* dan *token rarity* menegaskan bahwa kedua faktor tersebut mempengaruhi kinerja model, dengan BERT mempertahankan keunggulan tipis di seluruh kondisi pengujian. Secara praktis dari temuan ini menunjukkan bahwa untuk aplikasi yang sangat memprioritaskan akurasi—seperti pemrosesan teks pada domain legal atau medis, BERT tetap menjadi pilihan yang lebih andal. Sebaliknya, untuk sistem berskala besar, *real-time*, atau tertanam (*embedded*), MiniLM menawarkan keseimbangan efektif antara kecepatan dan akurasi.

Sejalan dengan prinsip *scaling laws*, peningkatan kapasitas model cenderung memberikan kinerja lebih baik, meskipun dengan konsekuensi kebutuhan komputasi yang meningkat secara signifikan.

Arah penelitian selanjutnya dapat mencakup pendekatan hibrida, seperti sistem yang menggunakan MiniLM untuk sebagian besar masukan, dan hanya mengalihkan ke BERT untuk kasus yang lebih kompleks.

DAFTAR REFERENSI

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[3] Google Research, "BERT: Bidirectional Encoder Representations from Transformers," GitHub repository. [Online]. Available: https://github.com/google-research/bert

[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint*, arXiv:1503.02531, 2015.

[5] W. Wang *et al*., "MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2140–2151, doi: 10.18653/v1/2021.findings-acl.188.

[6] Microsoft, "UniLM: MiniLM—State-of-the-art natural language processing," GitHub repository. [Online]. Available: https://github.com/microsoft/unilm/tree/master/minilm

[7] S. Mukherjee *et al*., "Orca: Progressive learning from complex explanation traces of GPT-4," *arXiv preprint*, arXiv:2306.02707, 2023.

[8] P. Zhang, G. Zeng, T. Wang, and W. Lu, "TinyLlama: An open-source small language model," *arXiv preprint*, arXiv:2401.02385, 2024.

[9] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT: A distilled version of BERT: Smaller, faster, cheaper and lighter," in *Proc. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (NeurIPS)*, 2019. [Online]. Available: arXiv:1910.01108.

[10] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: A compact task-agnostic BERT for resource-limited devices," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 2158–2170, doi: 10.18653/v1/2020.acl-main.195.

[11] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021, doi: 10.1007/s11263-021-01453-z.

[12] A. Wang *et al*., "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. NAACL-HLT 2018*, 2018, pp. 353–355, doi: 10.18653/v1/N18-2017.

[13] J.-T. Baillargeon and L. Lamontagne, "Assessing the impact of sequence length learning on classification tasks for transformer encoder models," *arXiv preprint*, arXiv:2212.08399, 2024.

[14] W. Yu *et al*., "Dict-BERT: Enhancing language model pre-training with dictionary," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1907–1918, doi: 10.18653/v1/2022.findings-acl.150.

Ken Ratri Retno Wardani, memperoleh gelar Magister dari Sekolah Teknik Elektro dan Informatika (STEI) Institut Teknologi Bandung. Bidang keahlian pengolahan citra dan kecerdasan buatan. Minat penelitian *computer vision* dan *deep learning*. Saat ini aktif mengajar di Program Studi Informatika Institut Teknologi Harapan Bangsa (ITHB) di Bandung.

Inge Martina, memperoleh gelar Magister dari Sekolah Teknik Elektro dan Informatika (STEI) di Institut Teknologi Bandung, bidang keahlian kecerdasan buatan. Minat penelitian salah satunya tentang *deep learning* dan *blockchain*. Saat ini aktif mengajar di Program Studi Informatika Institut Teknologi Harapan Bangsa (ITHB) di Bandung.

Jimmy Fong Xin Wern, kelahiran kota Bandung, saat ini tercatat sebagai mahasiswa semester akhir di Institut Teknologi Harapan Bangsa (ITHB) Bandung angkatan tahun 2021.