Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897

Hyperparameter Optimization Of IndoBERT Using Grid Search, Random Search, And Bayesian Optimization In Sentiment Analysis Of E-Government Application Reviews

Angga Iskoko[*1], Imam Tahyudin[2], Purwadi[3]

[1,2,3]Fakultas Ilmu Komputer, Universitas Amikom Purwokerto, Indonesia

Email: [1]anggaiskoko84@gmail.com

Received : Jun 17, 2025; Revised : Jul 3, 2025; Accepted : Jul 3, 2025; Published : Oct 16, 2025

Abstract

User reviews on Google Play Store reflect satisfaction and expectations regarding digital services, including E-Government applications. This study aims to optimize IndoBERT performance in sentiment classification through fine-tuning and hyperparameter exploration using three methods: Grid Search, Random Search, and Bayesian Optimization. Experiments were conducted on Sinaga Mobile app reviews, evaluated using accuracy, precision, recall, F1-score, learning curve, and confusion matrix. The results show that Grid Search with a learning rate of 5e-5 and a batch size of 16 provides the best results, with an accuracy of 90.55%, precision of 91.16%, recall of 90.55%, and F1-score of 89.75%. The learning curve indicates stable training without overfitting. This study provides practical contributions as a guide for improving IndoBERT in Indonesian sentiment analysis and as a foundation for developing NLP-based review monitoring systems to enhance public digital services.

Keywords : *Bayesian Optimization, E-Government, Grid Search, Hyperparameter, IndoBERT, Random Search.*

1. PENDAHULUAN

Dalam era transformasi digital, aplikasi berbasis *E-Government* menjadi salah satu sarana penting untuk mendukung penyelenggaraan pelayanan publik yang lebih efisien, transparan, dan akuntabel [1]. Pemerintah dituntut untuk terus menghadirkan layanan digital yang mudah diakses, responsif, serta sesuai dengan kebutuhan masyarakat . Salah satu indikator untuk menilai keberhasilan aplikasi tersebut adalah melalui ulasan pengguna (*app review*) yang dapat diakses secara terbuka di platform distribusi aplikasi, seperti Google Play Store [2]. Informasi yang terkandung dalam ulasan pengguna mencerminkan kepuasan, keluhan, saran, dan ekspektasi masyarakat terhadap performa dan kualitas layanan digital pemerintah [3]. Oleh karena itu, analisis sentimen terhadap ulasan pengguna memegang peranan vital untuk menghasilkan wawasan mendalam yang dapat dijadikan dasar evaluasi dan perbaikan berkelanjutan. Melalui hasil analisis sentimen yang akurat, pengembang aplikasi dan pihak pengambil kebijakan dapat merumuskan strategi pengembangan fitur maupun kebijakan pelayanan publik yang lebih tepat sasaran dan berorientasi pada kebutuhan pengguna.

Dalam konteks pemrosesan bahasa alami (*Natural Language Processing* / NLP) di Indonesia, IndoBERT sebuah model *Bidirectional Encoder Representations from Transformers* (BERT) yang dikembangkan khusus untuk bahasa Indonesia menjadi salah satu model berbasis *deep learning* yang populer digunakan untuk berbagai tugas NLP, termasuk klasifikasi teks dan analisis sentimen. Model ini memiliki keunggulan dalam memahami konteks kata dan hubungan antar frasa dalam satu kalimat berbahasa Indonesia [4]. Namun demikian, penerapan IndoBERT secara langsung tanpa penyesuaian parameter (*fine-tuning*) yang optimal sering kali belum menghasilkan performa klasifikasi sentimen yang maksimal, terutama pada data domain spesifik seperti ulasan aplikasi *E-Government* [5]. Oleh

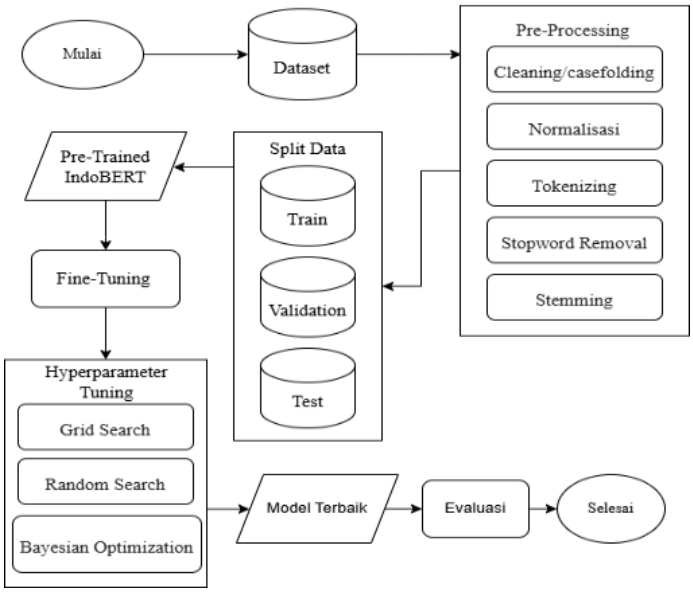
Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897

karena itu, diperlukan strategi optimasi hyperparameter yang terencana dan berbasis metode komputasi yang sistematis agar model dapat mencapai performa terbaiknya pada data target. Beberapa penelitian terdahulu menunjukkan bahwa pemilihan hyperparameter yang tepat dapat berpengaruh signifikan terhadap akurasi, kecepatan konvergensi, dan generalisasi model [6].

Meskipun IndoBERT telah banyak diadopsi dalam penelitian NLP di Indonesia, eksplorasi mendalam mengenai penerapan berbagai metode optimasi hyperparameter IndoBERT secara sistematis pada domain ulasan Google Play Store, khususnya aplikasi E-Government, masih jarang dijumpai. Misalnya, penelitian oleh Simanjuntak et al. (2024) hanya menerapkan hyperparameter tuning sederhana pada tugas deteksi berita palsu tanpa perbandingan metode optimasi yang sistematis. Sementara itu, Nugroho et al. (2020) melakukan fine-tuning BERT untuk analisis sentimen pada ulasan aplikasi, namun tidak membahas eksplorasi hyperparameter secara menyeluruh. Kedua penelitian tersebut belum mengevaluasi efektivitas berbagai pendekatan seperti Grid Search, Random Search, dan Bayesian Optimization dalam konteks aplikasi digital publik. Berdasarkan latar belakang dan permasalahan tersebut, penelitian ini bertujuan untuk mengevaluasi efektivitas beberapa metode hyperparameter tuning, yaitu Grid Search, Random Search, dan Bayesian Optimization, dalam proses *fine-tuning* IndoBERT untuk tugas analisis sentimen ulasan aplikasi *E-Government* di Google Play Store [9].

Novelty dari penelitian ini terletak pada eksplorasi dan perbandingan sistematis tiga pendekatan hyperparameter tuning yaitu grid Search, Random Search, dan Bayesian Optimization dalam konteks fine-tuning IndoBERT untuk analisis sentimen ulasan aplikasi E-Government di Google Play Store yang hingga saat ini belum banyak dibahas secra mendalam pada penelitian sebelumnya. Penelitian ini mengisi celah ilmiah dengan menghadirkan evaluasi komprehensif terhadap efektivitas metode optimasi tersebut pada data domain spesifik [8], sekaligus memberikan panduan praktis bagi pengembangan NLP berbahasa indonesia yang lebih adaptif dan presisi. Selain itu, penelitian ini juga diharapkan dapat mendukung upaya pemerintah dalam meningkatkan kualitas layanan publik digital yang berorientasi pada kepuasan dan kebutuhan masyarakat

2. METODE



Gambar 1. Metode Penelitian

Metode Penelitian dapat dilihat pada Gambar 1. Tahapan penelitian dimulai dengan pengumpulan dataset ulasan aplikasi sinaga mobile melalui scraping di Google Play Store pemrosesan data, kemudian dilanjutkan dengan tahap pre-processing, setelah itu data dibagi menjadi tiga bagian data

Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897

latih (training) 60%, data validasi 20%, dan data testing 20%. Model IndoBERT kemudian diterapkan melalui pre-trained IndoBERT dan fine-tuning. Selanjutnya untuk mendapatkan model terbaik dilakukan konfigurasi hyperparameter dengan berbagai metode seperti grid search, random search dan bayesian optimization. Terakhir model tersebut dievaluasi menggunakan metrik klasifikasi untuk menilai akurasi dan konsistensi prediksi sentimen oleh *IndoBERT*

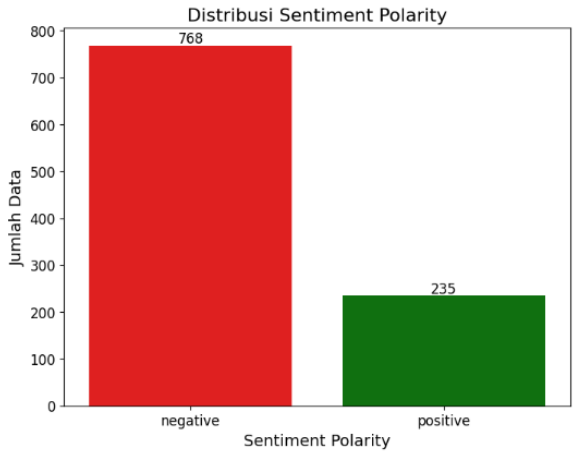
2.1. Dataset

Data yang digunakan dalam penelitian ini adalah ulasan pengguna dari aplikasi *Sinaga Mobile*, sebuah sistem informasi pelayanan kepegawaian digital yang dapat diunduh dari Google Play Store [10], dari periode waktu 6 Maret 2024 hingga 1 April 2025 sebanyak 1003 ulasan dengan metode scraping menggunakan pustaka pemrograman phyton bernama Google-Play-Scraper. Sebelum digunakan dalam proses pelatihan model, data melalui tahap *pre-processing* yang mencakup beberapa langkah, yaitu *cleaning* atau *casefolding* untuk mengubah huruf menjadi format seragam, *normalisasi* untuk menyamakan kata-kata tidak baku, *tokenizing* untuk memecah kalimat menjadi token, *stopword removal* untuk menghapus kata-kata umum yang tidak relevan, serta *stemming* untuk mengembalikan kata ke bentuk dasar [12]. Dalam penelitian ini, dataset diberi label menjadi dua kelas sentiment positive dan negative diambil dari 1003 ulasan aplikasi, pelabelan data dilakukan dengan metode *Lexicon Based* yang menggunakan *InSet Lexicon*, metode ini menghasilkan model klasifikasi. Dengan menggunakan kamus *lexicon* referensi yang didapat dari https://github.com/fajri91/InSet. Tabel 1. Menunjukkan hasil dari pelabelan data

Tabel 1. Hasil Pelabelan Data

Teks	Compound	Polarity
sering bug henti sendiri padahal android ram mohon baik sangat rugi konsumen	-13	negative
bug kadang nutup sendiri gagal absen mirip versi belum	-12	negative
lebih cepat merespon	4	positive
----------	-------	--------
sinaga dihp baru buka	2	positive

Hasil klasifikasi ditampilkan dalam bentuk distribusi sentimen, seperti yang terlihat pada Gambar 2, di mana mayoritas komentar dikategorikan sebagai negatif 768 ulasan, sedangkan komentar positif lebih sedikit 235 ulasan.



Gambar 2. Visualisasi Hasil Pelabelan

Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897

Dataset kemudian dibagi menjadi tiga bagian dengan proporsi 60% untuk data pelatihan (*training*), 20% untuk data validasi (*validation*), dan 20% untuk data pengujian (*testing*) [11]. Tabel 2 menunjukkan hasil split data.

Tabel 2 . Hasil Split Data

Train	601
Test	201
Validation	201

2.2. Model

Dalam penelitian ini model yang digunakan adalah IndoBERT, yaitu model *Bidirectional Encoder Representations from Transformers* (BERT) yang dirancang dan dilatih khusus untuk menangani teks berbahasa Indonesia [13]. *IndoBERT* memiliki keunggulan dalam memahami konteks kata, frasa, dan struktur kalimat dalam bahasa Indonesia secara lebih mendalam dibandingkan model BERT multibahasa [14]. Model pra-latih ini diunduh dari repositori *Hugging Face* yang sudah menyediakan Tansformers Library yang berisi ribuan model pra-latih (pretrained models) seperti BERT, GPT, RoBERTA, T5, IndoBERT dan lainnya. Pada penelitian ini model pra-latih yang digunakan adalah "*indobenchmark/indobert-base-p1*", yang telah melalui pelatihan awal menggunakan korpus teks Indonesia berskala besar. Kemudian, model dasar *IndoBERT* akan di-*fine-tune* lebih lanjut menggunakan dataset ulasan pengguna aplikasi *Sinaga Mobile* agar mampu melakukan klasifikasi sentimen secara lebih akurat pada domain ulasan aplikasi *E-Government* [15]. Dengan proses *fine-tuning* yang tepat, diharapkan *IndoBERT* dapat menyesuaikan bobot internalnya sesuai dengan karakteristik data target, sehingga menghasilkan prediksi sentimen yang lebih relevan dan handal

2.3. Fine Tuning IndoBERT

Proses *fine-tuning* dilakukan dengan menerapkan pengaturan parameter dasar yang bertujuan untuk menyesuaikan bobot model pra-latih IndoBERT agar lebih optimal dalam memahami pola dan karakteristik data ulasan aplikasi *Sinaga Mobile* [16]. Tahap ini merupakan langkah penting untuk meningkatkan akurasi model pada tugas klasifikasi sentimen yang spesifik, dengan tetap menjaga efisiensi komputasi [17]. Adapun nilai hyperparameter yang digunakan pada tahap *fine-tuning* awal ditetapkan berdasarkan rekomendasi praktik terbaik dalam pelatihan model BERT, sekaligus disesuaikan dengan kapasitas perangkat keras yang digunakan [18]. Nilai-nilai hyperparameter tersebut disajikan pada Tabel 3.

Tabel 3. Nilai Parameter Fine-tuning

Hyperparameter	Nilai
Learning Rate	2e-5
Batch Size	16
Epoch	10

2.4. Hyperparameter Tuning

Untuk mengoptimalkan performa *IndoBERT*, penelitian ini menggunakan tiga metode tuning hyperparameter, yaitu *Grid Search*, *Random Search*, dan *Bayesian Optimization*. *Grid Search* mengeksplorasi seluruh kombinasi parameter secara sistematis untuk menemukan konfigurasi terbaik [19]. *Random Search* memilih kombinasi secara acak, sehingga lebih efisien secara komputasi, terutama pada ruang parameter yang luas [20]. Sementara itu, *Bayesian Optimization* menggunakan pendekatan probabilistik dengan memanfaatkan hasil evaluasi sebelumnya, sehingga pencarian dilakukan secara

Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897

adaptif dan lebih hemat sumber daya [21]. Ketiga metode ini digunakan untuk membandingkan efektivitas strategi tuning dalam meningkatkan kinerja klasifikasi sentimen IndoBERT. Adapun rentang nilai hyperparameter yang diuji disajikan pada Tabel 4.

Tabel 4. Nilai Parameter *Hyperparameter Tuning*

Hyperparameter	Nilai
Learning Rate	2e-5, 3e-5, 5e-5
Batch Size	16, 32
Epoch	10

2.5. Evaluasi

Evaluasi performa model dilakukan menggunakan beberapa metrik klasifikasi untuk menilai akurasi dan konsistensi prediksi sentimen oleh *IndoBERT*. Metrik yang digunakan meliputi *Accuracy* (tingkat ketepatan keseluruhan), *Precision* (proporsi prediksi positif yang benar), Recall (kemampuan mendeteksi data positif), dan *F1-Score* (rata-rata harmonis precision dan recall), terutama penting untuk dataset dengan distribusi kelas tidak seimbang [22]. Nilai *Accuracy* didapatkan menggunakan persamaan (1), Nilai *Precision* didapatkan menggunakan persamaan (2), Nilai *Recall* didapatkan menggunakan persamaan (3), dan Nilai *F1-Score* didapatkan menggunakan persamaan (4).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

Selain itu, *Learning Curve* digunakan untuk memantau proses pembelajaran dan mendeteksi overfitting atau underfitting, sedangkan *Confusion Matrix* memberikan gambaran detail distribusi prediksi benar dan salah pada tiap kelas. Evaluasi dilakukan pada data validasi dan pengujian untuk mengukur kemampuan generalisasi model terhadap data baru [23].

3. HASIL

3.1. Hasil Fine Tuning IndoBERT

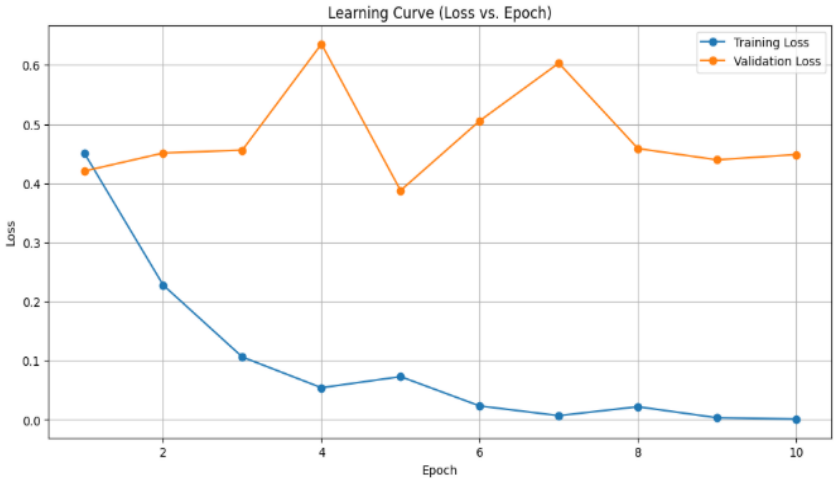
Hasil evaluasi pada Gambar 3, menunjukkan performa terbaik model IndoBERT setelah dilakukan proses fine-tuning dengan pengaturan parameter dasar. Model berhasil mencapai nilai accuracy sebesar 0.8955, yang berarti sekitar 89,55% data uji berhasil diprediksi dengan tepat. Nilai precision sebesar 0.8945 menunjukkan bahwa mayoritas prediksi positif yang dibuat oleh model adalah benar, sementara nilai recall sebesar 0.8955 mengindikasikan kemampuan model untuk mengidentifikasi data positif dengan tingkat keberhasilan yang tinggi. Sementara itu, nilai F1-Score sebesar 0.8893 menunjukkan keseimbangan yang baik antara precision dan recall, menunjukkan bahwa model memiliki kinerja yang stabil meskipun kemungkinan terdapat distribusi kelas yang tidak seimbang pada dataset. Secara keseluruhan, hasil ini membuktikan bahwa fine-tuning *IndoBERT* dengan pengaturan parameter dasar sudah mampu menghasilkan performa klasifikasi sentimen yang cukup baik dan dapat dijadikan sebagai titik acuan untuk tahap optimasi hyperparameter selanjutnya. Pada Gambar 3 adalah hasil model terbaik Fine Tuning *IndoBERT*.

Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897

```
Final Test Accuracy: 0.8955
Final Test Precision: 0.8945
Final Test Recall: 0.8955
Final Test F1 Score: 0.8893
```

Gambar 3. Hasil Terbaik Fine-Tuning IndoBERT

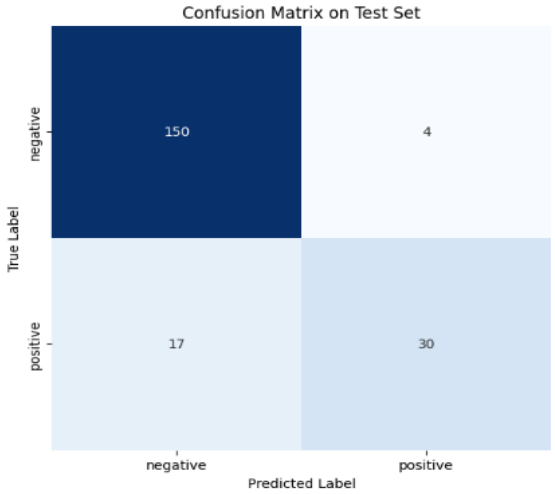
Selain nilai metrik evaluasi, performa model IndoBERT yang telah di-*fine-tune* juga dapat diamati melalui grafik learning curve pada Gambar 4. Grafik tersebut memperlihatkan pola perubahan *loss* pada data pelatihan (*training loss*) dan data validasi (*validation loss*) selama 10 epoch pelatihan. Terlihat bahwa nilai *training loss* mengalami penurunan yang konsisten dari epoch pertama hingga epoch terakhir, yang menunjukkan bahwa model memiliki kemampuan mempelajari pola data dengan baik pada tahap pelatihan. Sementara itu, *validation loss* tampak fluktuatif dengan beberapa kenaikan pada epoch tertentu, meskipun secara umum tetap berada pada rentang stabil di sekitar 0.4 hingga 0.6. Pola ini mengindikasikan bahwa model masih menjaga kemampuan generalisasi meskipun terdapat variasi pada data validasi. Secara keseluruhan, pola *learning curve* ini mendukung hasil evaluasi sebelumnya, di mana nilai *accuracy*, *precision*, *recall*, dan *F1-Score* yang tinggi memperlihatkan bahwa model IndoBERT yang di-*fine-tune* telah bekerja secara optimal tanpa indikasi *overfitting* yang signifikan. Dengan demikian, model ini layak dijadikan baseline untuk tahap optimasi hyperparameter lebih lanjut.



Gambar 4. Learning Curve Fine Tuning IndoBERT

Untuk melengkapi evaluasi performa model IndoBERT yang telah di-*fine-tune*, analisis lebih mendetail dilakukan melalui penyajian confusion matrix pada data pengujian. Gambaran distribusi prediksi model pada masing-masing kelas sentimen diberikan oleh Confusion Matrix, baik yang terklasifikasi dengan benar maupun yang salah klasifikasi. Berdasarkan confusion matrix pada Gambar 5, terlihat bahwa model mampu mengklasifikasikan ulasan dengan sentimen negatif secara sangat akurat, dengan 150 prediksi benar dan hanya 4 kesalahan prediksi sebagai positif. Sementara itu, untuk kelas sentimen positif, model memprediksi 30 sampel dengan benar dan terdapat 17 sampel yang salah terklasifikasi sebagai negatif. Hasil ini menunjukkan kemampuan deteksi model yang tinggi pada kelas negatif, namun masih perlu perbaikan pada prediksi kelas positif agar distribusi klasifikasi dapat lebih seimbang. Secara keseluruhan, confusion matrix ini mendukung interpretasi metrik evaluasi sebelumnya, di mana performa model IndoBERT hasil *fine-tuning* sudah memadai tetapi masih dapat dioptimalkan melalui tahap Tuning Hyperparameter.

Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897



Gambar 5. Hasil *Confusion Matrix Fine Tuning IndoBERT*

3.2. Hasil Hyperparameter Tuning Grid Search

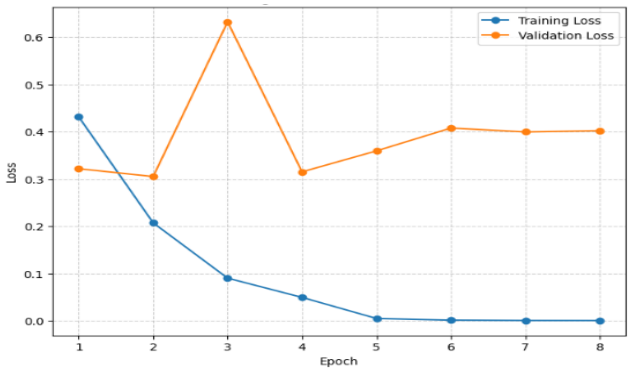
Setelah memperoleh baseline performa dari tahap *fine-tuning* awal, penelitian ini melanjutkan optimasi model dengan menerapkan metode hyperparameter tuning Grid Search. Grid Search digunakan untuk mengeksplorasi secara sistematis berbagai kombinasi hyperparameter guna menemukan konfigurasi terbaik yang dapat meningkatkan performa klasifikasi sentimen IndoBERT. Berdasarkan hasil tuning yang ditunjukkan pada Gambar 4, kombinasi hyperparameter terbaik yang diperoleh melalui Grid Search adalah learning rate sebesar 5e-05 dan batch size sebesar 16, dengan weight decay tetap 0.0. Dengan konfigurasi ini, model berhasil mencapai Test Accuracy sebesar 0.9055, Test Precision sebesar 0.9116, Test Recall sebesar 0.9055, dan Test F1-Score sebesar 0.8975. Peningkatan nilai metrik ini menunjukkan bahwa optimasi hyperparameter melalui Grid Search mampu memberikan kontribusi positif terhadap peningkatan akurasi dan keseimbangan kinerja model dibandingkan pengaturan parameter default pada tahap *fine-tuning* awal.

```
=== Best Model Result ===
Learning Rate: 5e-05
Weight Decay : 0.0
Batch Size   : 16
Test Accuracy: 0.9055
Test Precision: 0.9116
Test Recall: 0.9055
Test F1-score: 0.8975
```

Gambar 6. Hasil Terbaik Grid Search

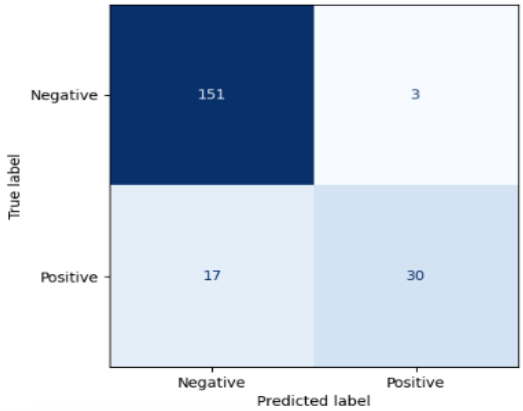
Sebagai pelengkap informasi performa model terbaik yang dihasilkan oleh Grid Search, Gambar 7, menampilkan learning curve yang merepresentasikan pola perubahan nilai *loss* pada data pelatihan (*training loss*) dan data validasi (*validation loss*) selama proses pelatihan berlangsung. Terlihat bahwa *training loss* mengalami penurunan tajam dan konsisten pada setiap epoch, yang menandakan model semakin mampu mempelajari pola data pelatihan dengan baik. Sementara itu, *validation loss* tampak relatif stabil meskipun terdapat fluktuasi pada beberapa epoch, dengan tren umum yang tetap terjaga di rentang 0.3 hingga 0.6. Pola ini menunjukkan bahwa model hasil Grid Search dapat menjaga keseimbangan antara pembelajaran pada data pelatihan dan kemampuan generalisasi pada data validasi. Hal ini selaras dengan peningkatan nilai metrik evaluasi sebelumnya, yang membuktikan bahwa optimasi hyperparameter melalui Grid Search berhasil menghasilkan model dengan performa yang lebih baik dan risiko *overfitting* yang rendah. Pada gambar 7 adalah tampilan learning curve Grid Search.

Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897



Gambar 7. *Learning Curve Grid Search*

Sebagai bagian dari evaluasi hasil optimasi hyperparameter menggunakan Grid Search, Gambar 8, menyajikan confusion matrix yang menggambarkan distribusi prediksi model pada data pengujian untuk masing-masing kelas sentimen. Berdasarkan confusion matrix tersebut, dapat dilihat bahwa model hasil Grid Search memiliki kemampuan prediksi yang sangat baik pada kelas sentimen negatif, dengan total 151 prediksi benar dan hanya 3 sampel yang salah diklasifikasikan sebagai positif. Sementara itu, pada kelas sentimen positif, model berhasil mengklasifikasikan 30 sampel dengan benar, meskipun masih terdapat 17 sampel positif yang salah terprediksi sebagai negatif. Jika dibandingkan dengan confusion matrix pada tahap *fine-tuning* awal, hasil ini menunjukkan adanya perbaikan pada prediksi kelas negatif dengan penurunan jumlah kesalahan prediksi. Secara keseluruhan, confusion matrix ini mendukung peningkatan nilai metrik evaluasi yang telah dicapai, serta memperlihatkan bahwa Grid Search efektif dalam memperbaiki akurasi dan keseimbangan klasifikasi antar kelas.



Gambar 8. *Confusion Matrix Grid Search*

3.3. Hasil *Random Search*

Setelah melakukan pencarian hyperparameter secara sistematis dengan *Grid Search*, penelitian ini juga menerapkan metode *Random Search* sebagai alternatif pendekatan optimasi. *Random Search* bekerja dengan cara melakukan sampling acak dari ruang hyperparameter yang telah ditetapkan, sehingga memungkinkan ditemukannya kombinasi parameter yang optimal dengan biaya komputasi yang lebih efisien, terutama pada ruang parameter yang luas. Berdasarkan hasil *Random Search* yang ditunjukkan pada Gambar 9, diperoleh konfigurasi hyperparameter terbaik dengan learning rate sebesar 3e-05, batch size sebesar 16, dan *weight decay* sebesar 0.0001. Dengan konfigurasi ini, model *IndoBERT* berhasil mencapai Test *Accuracy* sebesar 0.9005, Test Precision sebesar 0.8983, Test Recall sebesar 0.9005, dan Test *F1-Score* sebesar 0.8961. Hasil ini menunjukkan bahwa *Random Search* juga efektif dalam meningkatkan kinerja klasifikasi sentimen, meskipun nilai metriknya sedikit lebih rendah

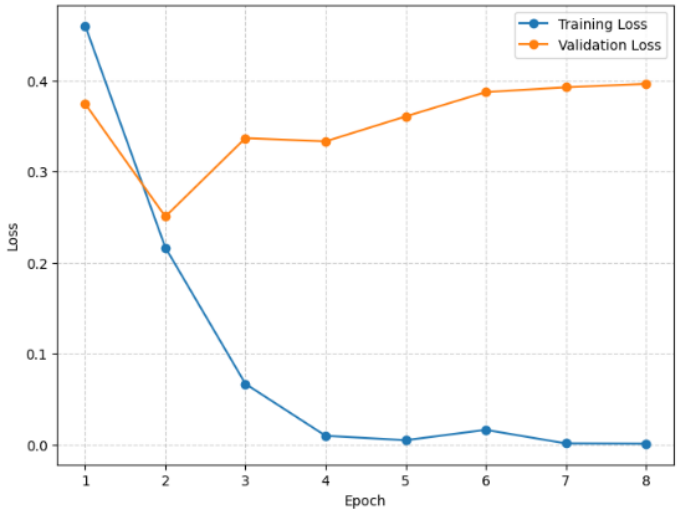
Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897

dibandingkan dengan hasil terbaik yang diperoleh melalui *Grid Search*. Secara keseluruhan, *Random Search* memberikan alternatif solusi optimasi yang cepat dengan hasil yang kompetitif.

```
========== Best Result ==========
Trial        : 1
Learning Rate: 3e-05
Batch Size   : 16
Weight Decay : 0.0001
Test Accuracy: 0.9005
Test Precision: 0.8983
Test Recall  : 0.9005
Test F1-score: 0.8961
```

Gambar 9. Hasil Terbaik *Random Search*

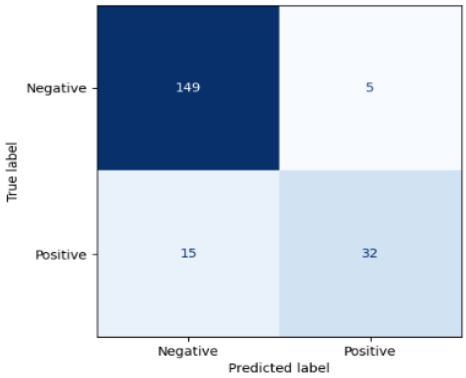
Gambar 10, menunjukkan grafik *Learning Curve* terbaik yang diperoleh dari hasil proses *Random Search* dalam optimasi hyperparameter pada model yang digunakan. Berdasarkan hasil pencarian, kombinasi hyperparameter yang memberikan performa terbaik adalah *Learning Rate* sebesar 3e-05 dan *Batch Size* sebanyak 16. Pada konfigurasi ini, terlihat bahwa *Training Loss* mengalami penurunan tajam dan konsisten hingga mendekati nol, yang menandakan bahwa model mampu belajar dengan sangat baik terhadap data pelatihan. Namun, *Validation Loss* justru menunjukkan tren peningkatan setelah beberapa epoch awal, yang mengindikasikan adanya gejala *overfitting*. Meski demikian, grafik ini tetap menjadi acuan penting dalam mengevaluasi performa model serta menentukan langkah selanjutnya untuk perbaikan, seperti regularisasi atau *early stopping*.



Gambar 10. *Learning Curve Random Search*

Gambar 11, menunjukkan *Confusion Matrix* dari hasil evaluasi model terbaik yang diperoleh melalui pendekatan Random Search dalam proses tuning hyperparameter. Pada hasil ini, model mampu mengklasifikasikan komentar dengan performa yang cukup baik. Tercatat sebanyak 149 data berlabel *Negative* berhasil diprediksi dengan benar (*True Negative*), sedangkan 32 data *Positive* juga diklasifikasikan secara tepat (*True Positive*). Meski demikian, masih terdapat 5 data negatif yang salah diklasifikasikan sebagai positif (*False Positive*) dan 15 data positif yang salah diklasifikasikan sebagai negatif (*False Negative*). Matriks ini mencerminkan bahwa model hasil Random Search memiliki kecenderungan yang kuat dalam mengenali kelas negatif, namun masih memiliki keterbatasan dalam mengenali kelas positif secara optimal.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897



Gambar 11. Confusion Matrix Random Search

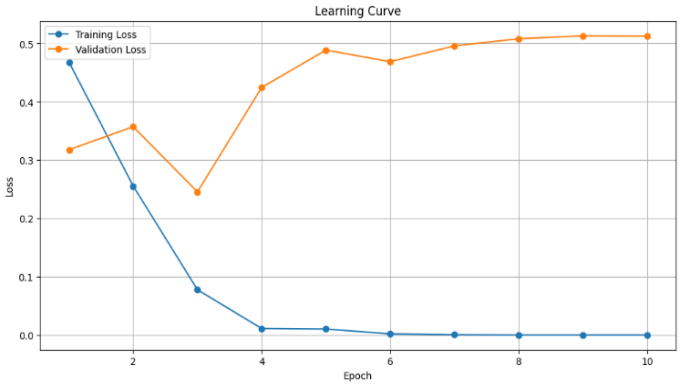
3.4. Hasil Bayesian Optimization

Berdasarkan hasil Bayesian Optimization, diperoleh kombinasi hyperparameter yang menghasilkan performa optimal dengan nilai Test Accuracy sebesar 0.8905, Test Precision sebesar 0.8896, Test Recall sebesar 0.8905, dan Test F1-Score sebesar 0.8833. Meskipun nilai-nilai metrik ini sedikit lebih rendah dibandingkan hasil dari Grid Search, performa yang dicapai tetap kompetitif dan menunjukkan kestabilan model yang baik. Selain itu, pendekatan Bayesian Optimization menawarkan keunggulan dalam efisiensi waktu dan sumber daya komputasi, karena tidak perlu mengevaluasi semua kombinasi secara eksplisit seperti pada Grid Search. Gambar 12 di bawah ini memperlihatkan ringkasan metrik evaluasi terbaik yang diperoleh dari hasil tuning hyperparameter menggunakan Bayesian Optimization.

```
--- Best Metrics Summary ---
Best Test Accuracy : 0.8905
Best Test Precision: 0.8896
Best Test Recall   : 0.8905
Best Test F1-score : 0.8833
```

Gambar 12. Hasil Terbaik Bayesian Optimization

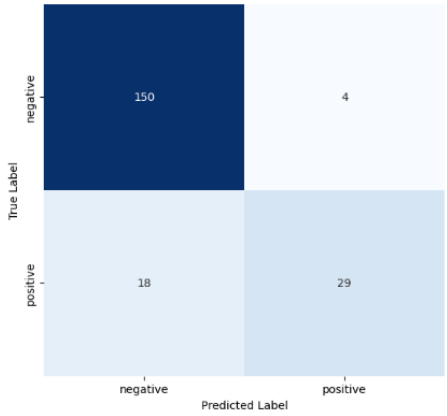
Learning Curve terbaik pada hasil Bayesian Optimization diperoleh dengan kombinasi hyperparameter learning rate sebesar 5e-05 dan batch size 16. Grafik menunjukkan bahwa training loss menurun tajam hingga mendekati nol, menandakan model belajar dengan sangat baik terhadap data pelatihan. Namun, validation loss cenderung meningkat setelah epoch ke-4, yang mengindikasikan terjadinya overfitting. Meskipun begitu, performa model tetap kompetitif dan stabil berdasarkan evaluasi metrik lainnya. Gambar 13 di bawah ini memperlihatkan hasil learning curve menggunakan Bayesian Optimization.



Gambar 13. Learning Curve Bayesian Optimization

Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897

Untuk melengkapi evaluasi performa model dari hasil Bayesian Optimization pada Gambar 14, menampilkan confusion matrix yang menunjukkan distribusi prediksi model pada data pengujian. Model berhasil mengklasifikasikan 150 ulasan negatif dengan benar, dan hanya 4 ulasan negatif yang salah diklasifikasikan sebagai positif. Sementara itu, pada kelas positif, terdapat 29 prediksi yang benar, dan 18 ulasan positif yang salah diklasifikasikan sebagai negatif. Hasil ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengenali sentimen negatif, namun masih perlu peningkatan dalam mendeteksi sentimen positif secara lebih akurat. Meskipun begitu, distribusi prediksi ini tetap mencerminkan kinerja model yang kompetitif dan seimbang secara umum



Gambar 14. *Confusion Matrix Bayesian Optimization*

3.5. Perbandingan model

Untuk memberikan gambaran menyeluruh mengenai performa masing-masing metode optimasi yang diterapkan. Tabel 5, menyajikan ringkasan nilai metrik evaluasi utama (Accuracy, Precision, Recall, dan F1-Score) dari setiap pendekatan yang digunakan dalam penelitian ini.

Tabel 5. Perbadingan Model yang Digunakan

Metode	Learning Rate	Batch Size	Accuracy	Precision	Recall	F1-Score
Fine-Tuning Awal	2e-5	16	0.8955	0.8945	0.8955	0.8893
Grid Search (Terbaik)	5e-5	16	0.9055	0.9116	0.9055	0.8975
Random Search	3e-5	16	0.9005	0.8983	0.9005	0.8961
Bayesian Optimization	5e-5	16	0.8905	0.8896	0.8905	0.8833

Berdasarkan Tabel 5, metode Grid Search menunjukkan performa terbaik dibandingkan pendekatan lain. Dengan konfigurasi learning rate 5e-5 dan batch size 16, model mencapai accuracy 90.55%, precision 91.16%, recall 90.55%, dan F1-score 89.75%. Keunggulan Grid Search terletak pada eksplorasi sistematis terhadap kombinasi hyperparameter, meskipun membutuhkan waktu komputasi lebih tinggi. Ini membuktikan efektivitas tuning terstruktur dalam meningkatkan performa *IndoBERT*. Sementara itu, Random Search menawarkan efisiensi waktu dengan hasil yang kompetitif, meski menunjukkan gejala overfitting. Bayesian Optimization unggul dalam efisiensi proses, namun performanya masih di bawah dua metode lainnya. Secara keseluruhan, Grid Search dinilai sebagai metode paling efektif untuk mengoptimalkan *IndoBERT* pada analisis sentimen ulasan aplikasi *E-Government*.

Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897

4. DISKUSI

4.1. Keunggulan Metode

Setiap metode tuning yang digunakan memiliki keunggulan masing-masing dalam meningkatkan performa IndoBERT. Grid Search terbukti paling optimal dengan performa tertinggi pada seluruh metrik evaluasi. Pendekatannya yang sistematis menjelajahi seluruh kombinasi parameter, meskipun memerlukan waktu komputasi lebih besar, memastikan tidak ada konfigurasi terbaik yang terlewat.

Random Search unggul dalam efisiensi waktu karena hanya mengambil sampel acak dari ruang parameter. Hasilnya sedikit di bawah Grid Search, namun tetap kompetitif dan cocok digunakan saat sumber daya terbatas. Meski demikian, grafik learning curve menunjukkan adanya indikasi overfitting, yang menandakan model terlalu menyesuaikan diri dengan data pelatihan.

Bayesian Optimization menggunakan pencarian adaptif berbasis probabilistik yang efisien dalam jumlah eksperimen. Meskipun metriknya tidak setinggi Grid Search, metode ini tetap layak dipertimbangkan, khususnya pada eksperimen berskala besar dengan keterbatasan waktu atau perangkat keras.

Tuning hyperparameter memberikan dampak signifikan terhadap performa model. Nilai learning rate yang terlalu besar dapat membuat model gagal konvergen, sedangkan nilai terlalu kecil memperlambat proses belajar. Dalam penelitian ini, kombinasi learning rate 5e-5 dan batch size 16 menghasilkan akurasi tertinggi sebesar 90.55%.

Batch size juga memengaruhi kestabilan pelatihan. Ukuran kecil memberikan pembaruan bobot yang lebih halus, sedangkan ukuran besar mempercepat pelatihan namun berisiko kehilangan variasi data. Kombinasi parameter yang tepat membantu model belajar efisien dan mencegah underfitting maupun overfitting. Secara keseluruhan, strategi tuning yang tepat sangat penting dalam fine-tuning IndoBERT, terutama pada domain spesifik seperti ulasan layanan publik, dan merupakan bagian krusial dari pipeline NLP berbasis deep learning.

Penelitian ini memiliki keunggulan dalam pendekatan komparatif yang menyeluruh terhadap tiga metode hyperparameter tuning. Penggunaan model IndoBERT yang di-fine-tune untuk ulasan aplikasi E-Government memberikan kontribusi signifikan dalam konteks lokal dan bahasa Indonesia [24].

Namun, terdapat beberapa keterbatasan. Data yang digunakan hanya dari satu aplikasi, yaitu *Sinaga Mobile*, sehingga generalisasi ke aplikasi lain masih terbatas. Selain itu, eksplorasi hyperparameter hanya mencakup learning rate dan batch size, belum mencakup parameter penting lain seperti weight decay, dropout, dan max sequence length [25]. Belum diterapkannya strategi regularisasi seperti early stopping juga menyebabkan potensi overfitting pada beberapa konfigurasi.

4.2. Dampak Overfitting

Hasil analisis grafik learning curve menunjukkan indikasi overfitting pada model *IndoBERT*, ditandai dengan training loss yang terus menurun hingga mendekati nol, sementara validation loss meningkat setelah beberapa epoch awal [26]. Hal ini menunjukkan bahwa model terlalu menyesuaikan diri dengan data pelatihan sehingga tidak dapat generalisasi dengan data baru. Salah satu penyebab utamanya adalah kompleksitas arsitektur *IndoBERT* yang memiliki sekitar 110 juta parameter, membuatnya rentan overfitting pada dataset yang terbatas. Selain itu, jumlah data dan variasi yang kurang optimal dalam dataset ulasan *Sinaga Mobile* membuat model cenderung menghafal detail spesifik daripada mempelajari representasi yang dapat digeneralisasi [27]. Faktor lain adalah pengaturan hyperparameter, seperti penggunaan epoch yang terlalu banyak tanpa early stopping, nilai weight decay yang rendah atau tidak digunakan sama sekali, serta ketiadaan dropout, yang semuanya meningkatkan risiko overfitting [28]. Di sisi lain, learning rate yang cukup tinggi seperti 3e-5 memang mempercepat konvergensi, namun dapat memperparah overfitting jika tidak dibarengi teknik regularisasi. Oleh karena

itu, kombinasi pengaturan parameter yang tepat serta penerapan strategi regularisasi seperti weight decay dan early stopping sangat penting untuk menjaga kemampuan generalisasi model.

4.3. Rekomendasi

Berdasarkan temuan dalam penelitian ini, disarankan agar penelitian lanjutan menerapkan berbagai strategi untuk memitigasi overfitting dan meningkatkan kemampuan generalisasi model *IndoBERT*. Salah satu langkah yang penting adalah menerapkan early stopping dengan pemantauan ketat terhadap metrik validasi seperti validation loss atau F1-score, agar pelatihan dapat dihentikan secara otomatis saat performa mulai menurun. Selain itu, penambahan dropout layer pada bagian classifier head *IndoBERT* dapat membantu mencegah model terlalu bergantung pada fitur tertentu. Penggunaan weight decay dengan nilai yang lebih besar, seperti 0.01, juga disarankan untuk membatasi kompleksitas model. Untuk mengatasi keterbatasan variasi data, dapat dilakukan augmentasi data teks, seperti synonym replacement atau back-translation, guna memperkaya distribusi data pelatihan. Terakhir, memperbesar ukuran dataset akan sangat membantu dalam meningkatkan kapasitas generalisasi model terhadap data baru. Dengan menerapkan strategi-strategi ini, diharapkan performa IndoBERT tidak hanya optimal pada data pelatihan, tetapi juga stabil pada data validasi dan pengujian.

5. KESIMPULAN

Penelitian ini menyimpulkan bahwa proses hyperparameter tuning memberikan dampak signifikan terhadap performa model *IndoBERT* dalam klasifikasi sentimen ulasan aplikasi E-Government. Dari tiga metode yang diuji, *Grid Search* menghasilkan performa terbaik, temuan ini menunjukkan bahwa pendekatan tuning yang sistematis mampu meningkatkan akurasi dan stabilitas model secara signifikan. Implikasi praktis dari penelitian ini dapat diterapkan dalam pengembangan sistem pemantauan ulasan otomatis, khususnya untuk mengawasi tanggapan publik terhadap aplikasi layanan digital pemerintah melalui Google Play Store. Model yang telah dioptimalkan dapat digunakan untuk mengidentifikasi keluhan, saran, maupun apresiasi pengguna secara real-time, sehingga dapat menjadi dasar dalam pengambilan keputusan berbasis data. Untuk pengembangan lanjutan, disarankan memperluas dataset ke berbagai aplikasi E-Government guna meningkatkan generalisasi model, serta menerapkan strategi regularisasi seperti early stopping, dropout, dan weight decay untuk mencegah overfitting. Eksplorasi terhadap model transformer lain seperti *IndoBERTweet* atau *RoBERTa* Indonesia juga dapat dipertimbangkan, termasuk integrasi teknik augmentasi data dan pendekatan *aspect-based sentiment analysis* (ABSA) untuk menghasilkan analisis sentimen yang lebih mendalam dan kontekstual. Penelitian ini memberikan kontribusi ilmiah yang nyata dalam penguatan ekosistem NLP berhasa indonesia, khususnya melalui pembuktian efektivitas fine-tuning dan optimasi hyperparameter pada model IndoBERT dalam domain aplikasi e-government.

DAFTAR PUSTAKA

[1] R. Artikel, M. I. Amal, E. S. Rahmasita, E. Suryaputra, and N. A. Rakhmawati, “Analisis Klasifikasi Sentimen Terhadap Isu Kebocoran Data Kartu Identitas Ponsel di Twitter Sentiment Classification Analysis On Phone Identity Card Data Leaks Issues On Twitter,” vol. 8, no. September, pp. 645–660, 2022.

[2] K. A. Pradani, L. H. Suadaa, J. Timur, and P. Korespondensi, “Automated Essay Scoring Menggunakan Semantic Textual Automated Essay Scoring Using Transformer-Based Semantic,” vol. 10, no. 6, pp. 1177–1184, 2023, doi: 10.25126/jtiik.2023107338.

[3] G. Z. Nabiilah, S. Y. Prasetyo, Z. N. Izdihar, and A. S. Girsang, “ScienceDirect ScienceDirect 7th International Conference on Computer Science and Computational Intelligence 2022 BERT base model for toxic comment analysis on BERT base model for toxic comment analysis on Indonesian social media Indonesian social media,” *Procedia Comput. Sci.*, vol. 216, no. 2022, pp. 714–721, 2023, doi: 10.1016/j.procs.2022.12.188.

Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897

[4] J. Nasional, S. Informasi, E. Teks, E. Marthen, and I. Novita, "Optimasi RoBERTa dengan Hyperparameter Tuning untuk Deteksi," vol. 03, pp. 240–248, 2024.
[5] A. Simanjuntak, R. Lumbantoruan, K. Sianipar, R. Gultom, M. Simaremare, and S. Situmeang, "Studi dan Analisis Hyperparameter Tuning IndoBERT Dalam Pendeteksian Berita Palsu," vol. 13, pp. 60–67, 2024.
[6] K. S. Nugroho, A. Y. Sukmadewa, F. A. Bachtiar, and N. Yudistira, "BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews," pp. 1–10, 2020.
[7] X. Teng, L. Zhang, P. Gao, C. Yu, and S. Sun, "BERT-Driven stock price trend prediction utilizing tokenized stock data and multi-step optimization approach," *Appl. Soft Comput.*, vol. 170, no. December 2024, p. 112627, 2025, doi: 10.1016/j.asoc.2024.112627.
[8] P. Nur *et al.*, "Bijakaweb : Platform Berbasis Web Untuk Deteksi Hate Speech Bijakaweb : A Web-Based Platform For Detecting Hate Speech In," vol. 11, no. 4, pp. 939–948, 2024, doi: 10.25126/jtiik.1148719.
[9] A. Turchin, S. Masharsky, and M. Zitnik, "Informatics in Medicine Unlocked Comparison of BERT implementations for natural language processing of narrative medical documents," *Informatics Med. Unlocked*, vol. 36, no. November 2022, p. 101139, 2023, doi: 10.1016/j.imu.2022.101139.
[10] T. D. Purnomo, J. Sutopo, and A. History, "Comparison Of Pre-Trained Bert-Based Transformer Models For Regional," vol. 3, no. 3, pp. 11–21, 2024.
[11] K. Kunci, "The Indonesian Journal of Computer Science," vol. 13, no. 5, pp. 8350–8359, 2024.
[12] H. S. Anggraheni, M. J. Naufal, and N. Yudistira, "Deteksi Spam Berbahasa Indonesia Berbasis Teks Menggunakan Model Bert Text-Based Indonesian Spam Detection Using The Bert Model," vol. 11, no. 6, pp. 1291–1301, 2024, doi: 10.25126/jtiik.2024118121.
[13] R. I. Perwira and V. A. Permadi, "Domain-Specific Fine-Tuning of IndoBERT for Aspect- Based Sentiment Analysis in Indonesian Travel User- Generated Content," vol. 11, no. 1, pp. 30–40, 2025.
[14] A. Alamsyah and Y. Sagama, "Intelligent Systems with Applications Empowering Indonesian internet users : An approach to counter online toxicity and enhance digital well-being," *Intell. Syst. with Appl.*, vol. 22, no. August 2023, p. 200394, 2024, doi: 10.1016/j.iswa.2024.200394.
[15] F. Baharuddin, "Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom ' s Taxonomy," vol. 9, no. 2, 2023.
[16] I. Griha, T. Isa, and P. N. Sriwijaya, "Hyperparameter Tuning Epoch dalam Meningkatkan Akurasi Data Latih dan Data Validasi pada Citra Pengendara," no. November 2022, 2023, doi: 10.36499/psnst.v12i1.6697.
[17] M. Evtimova, "Hyperparameter Tuning for Address Validation using Optuna Applied for Address Validation 2 Related Work 3 Standards for Postal Address in France," vol. 12, pp. 105–111, 2024, doi: 10.37394/232018.2024.12.10.
[18] T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," 2020.
[19] Z. A. Annisa, R. S. Perdana, P. P. Adikara, U. Brawijaya, and P. Korespondensi, "Kombinasi Intent Classification Dan Named Entity Recognition Pada Data Berbahasa Indonesia Dengan Metode Dual Intent And Combining Intent Classification And Named Entity Recognition On," vol. 11, no. 5, pp. 1017–1024, 2024, doi: 10.25126/jtiik.2024117985.
[20] M. N. Zaidan, Y. Sibaroni, and S. S. Prasetiyowati, "LEARNING RATE AND EPOCH OPTIMIZATION IN THE FINE-TUNING PROCESS FOR INDO BERT ' S PERFORMANCE ON SENTIMENT ANALYSIS OF," vol. 5, no. 5, pp. 1443–1450, 2024.
[21] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh Tahapan Preprocessing Terhadap Model Indobert Dan Indobertweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 4, pp. 887–894, 2024, doi: 10.25126/jtiik.1148315.
[22] I. D. A. N. I. Pada, "Perbandingan Kinerja Pre-Trained Tokopedia Seller Center," vol. 11, no. 2, pp. 13–20, 2024, doi: 10.30656/jsii.v11i2.9168.
[23] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 13, no. 2, pp. 1–43,

Vol. 6, No. 5, October 2025, Page. 3430-3444
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4897

2023, doi: 10.1002/widm.1484.

[24] R. Pramana, M. Jonathan, H. S. Yani, and R. Sutoyo, "A Comparison of BiLSTM, BERT, and Ensemble Method for Emotion Recognition on Indonesian Product Reviews," *Procedia Comput. Sci.*, vol. 245, no. C, pp. 399–408, 2024, doi: 10.1016/j.procs.2024.10.266.

[25] R. Nugroho, N. Azka, W. Sayudha, and R. Graha, "Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi) Analisis Sentimen Ulasan Aplikasi Mobile JKN di Google," vol. 9, no. June, pp. 495–505, 2025.

[26] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 9, no. 3, pp. 746–757, 2023, doi: 10.26555/jiteki.v9i3.26490.

[27] P. Jeevallucas et al., "Analisis Sentimen Pengguna terhadap Akun X/Twitter Resmi 'DANA' dengan Algoritma IndoBERT," Jurnal Teknologi Informasi dan Komunikasi, vol. 8, no. 11, pp. 549–559, 2025.

[28] C. J. L. Tobing, I. G. N. L. Wijayakusuma, L. Putu, I. Harini, and U. Udayana, "Detection of Political Hoax News Using Fine-Tuning IndoBERT," vol. 9, no. 2, pp. 354–360, 2025.