JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9545
Hal 221-233
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Optimasi IndoBERT untuk Pengenalan Entitas Bernama Bahasa Indonesia pada Data Media Sosial dengan Penalaan Hiperparameter Optuna

Bambang Siswanto*, M. Hanafi

Fakultas Ilmu Komputer, PJJ Informatrika, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
Email: [1*]basisjoss@students.amikom.ac.id, [2]Hanafi@amikom.ac.id
Email Penulis Korespondensi: basisjoss@students.amikom.ac.id
Submitted **27-01-2026;** Accepted **12-02-2026;** Published **28-02-2026**

Abstrak

Named Entity Recognition (NER) atau pengenalan entitas bernama merupakan tugas fundamental dalam pemrosesan bahasa alami untuk mengekstraksi informasi terstruktur dari teks tidak terstruktur. Pada Bahasa Indonesia, khususnya teks media sosial yang informal dan bervariasi, kinerja model NER berbasis Bidirectional Encoder Representations from Transformers (BERT) sangat dipengaruhi oleh konfigurasi hiperparameter pada tahap fine-tuning. Namun, banyak studi masih menggunakan konfigurasi bawaan atau penyesuaian terbatas sehingga potensi peningkatan kinerja dan stabilitas belum sepenuhnya dimanfaatkan. Penelitian ini mengevaluasi dampak penalaan hiperparameter menggunakan Optuna berbasis Tree-structured Parzen Estimator (TPE) terhadap kinerja dan stabilitas pelatihan IndoBERT (indobenchmark/indobert-base-p1) pada data Twitter/X. Kontribusi utama penelitian ini adalah menyajikan evaluasi empiris mengenai pengaruh penalaan hiperparameter terhadap peningkatan kinerja dan stabilitas pelatihan IndoBERT, serta menghasilkan rekomendasi konfigurasi yang andal untuk replikasi dan penerapan NER Bahasa Indonesia. Dataset dianotasi dengan skema pelabelan Begin–Inside–Outside (BIO) untuk tiga tipe entitas, yaitu person (PER), organization (ORG), dan location (LOC). Fungsi objektif optimasi ditetapkan sebagai F1-score pada data validasi. Hasil menunjukkan konfigurasi Optuna mencapai precision 0,9338, recall 0,9312, F1-score 0,9325, dan accuracy 0,9854 pada data uji, melampaui baseline dengan F1-score 0,9253 dan accuracy 0,9837. Evaluasi multi-seed menunjukkan peningkatan yang konsisten, dengan F1 rata-rata 0,9302 ± 0,0016 dibandingkan 0,9238 ± 0,0009 pada baseline. Temuan ini menegaskan bahwa penalaan hiperparameter berbasis Optuna meningkatkan kinerja sekaligus keandalan IndoBERT untuk NER Bahasa Indonesia pada teks media sosial..

Kata Kunci: Named Entity Recognition; IndoBERT; Bahasa Indonesia; Optimasi Hiperparameter; Optuna

Abstract

Named Entity Recognition (NER) is a fundamental task in natural language processing for extracting structured information from unstructured text. In Indonesian, particularly for informal and diverse social media text, the performance of NER models based on Bidirectional Encoder Representations from Transformers (BERT) is strongly influenced by hyperparameter configurations during fine-tuning. However, many studies still rely on default settings or limited adjustments, so the potential improvements in performance and training stability have not been fully exploited. This study evaluates the impact of hyperparameter tuning using Optuna with a Tree-structured Parzen Estimator (TPE) on the performance and training stability of IndoBERT (indobenchmark/indobert-base-p1) on Twitter/X data. The main contribution of this work is an empirical evaluation of how hyperparameter tuning improves IndoBERT's performance and training stability, and the resulting recommendations of reliable configurations for reproducible experiments and practical deployment of Indonesian NER. The dataset is annotated using the Begin–Inside–Outside (BIO) labeling scheme for three entity types: person (PER), organization (ORG), and location (LOC). The optimization objective is defined as the F1-score on the validation set. The results show that the Optuna configuration achieves a precision of 0.9338, recall of 0.9312, F1-score of 0.9325, and accuracy of 0.9854 on the test set, outperforming the baseline with an F1-score of 0.9253 and accuracy of 0.9837. Multi-seed evaluation indicates consistent improvements, with an average F1 of 0.9302 ± 0.0016 compared to 0.9238 ± 0.0009 for the baseline. These findings confirm that Optuna-based hyperparameter tuning improves both the performance and reliability of IndoBERT for Indonesian NER on social media text.

Keywords: Named Entity Recognition; IndoBERT; Indonesian Language; Hyperparameter Optimization; Optuna

1. PENDAHULUAN

Named Entity Recognition (NER) merupakan salah satu tugas inti dalam pemrosesan bahasa alami (*Natural Language Processing/NLP*) yang berfokus pada identifikasi dan klasifikasi entitas bernama, seperti *Person* (PER), *Organization* (ORG), dan *Location* (LOC), dari teks tidak terstruktur. Keberhasilan NER sangat menentukan kualitas berbagai aplikasi NLP lanjutan, termasuk ekstraksi informasi, sistem penjawab pertanyaan, dan analisis dokumen. Namun, pada Bahasa Indonesia, kinerja NER masih menghadapi tantangan signifikan karena keterbatasan sumber daya linguistik serta tingginya ketergantungan model terhadap konfigurasi pelatihan yang digunakan [1],[2]. Secara khusus, model NER berbasis Transformer sering menunjukkan variasi performa yang besar ketika konfigurasi hiperparameter *fine-tuning* tidak ditentukan secara tepat.

Perkembangan arsitektur Transformer, khususnya BERT (*Bidirectional Encoder Representations from Transformers*), telah membawa peningkatan signifikan pada berbagai tugas NLP, termasuk NER [3]. Untuk Bahasa Indonesia, IndoBERT dikembangkan sebagai model pralatih monolingual yang dilatih menggunakan korpus besar berbahasa Indonesia dan terbukti unggul dibandingkan model general dalam berbagai tugas NLP [4][5]. Meskipun demikian, keunggulan IndoBERT tidak secara otomatis menjamin performa optimal pada tugas NER. Proses *fine-tuning* tetap menjadi faktor penentu utama, terutama dalam penyesuaian Hiperparameter seperti *learning rate*, *batch size*, *weight decay*, dan *warmup ratio*, yang berpengaruh langsung terhadap konvergensi dan generalisasi model [6].

This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9545
Hal 221-233
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Sejumlah penelitian menunjukkan bahwa pemilihan hiperparameter yang tidak tepat dapat menyebabkan pelatihan model menjadi tidak stabil, menghasilkan fluktuasi performa yang tinggi antar percobaan meskipun menggunakan data dan arsitektur yang sama [7]. Dalam praktiknya, banyak penelitian NER masih mengandalkan konfigurasi hiperparameter default atau mengadopsi konfigurasi dari penelitian sebelumnya tanpa analisis empiris yang memadai. Pendekatan ini berpotensi menghasilkan model yang suboptimal dan sulit direproduksi, terutama pada bahasa dengan karakteristik linguistik yang berbeda seperti Bahasa Indonesia.

Untuk mengatasi permasalahan tersebut, berbagai pendekatan optimasi hiperparameter telah dikembangkan. Metode konvensional seperti *grid search* dan *random search* sering digunakan, namun memiliki keterbatasan dalam efisiensi eksplorasi ruang pencarian ketika jumlah hiperparameter meningkat. Pendekatan optimasi hiperparameter berbasis Bayesian menawarkan solusi yang lebih adaptif dengan memanfaatkan informasi dari percobaan sebelumnya. Optuna merupakan salah satu kerangka kerja optimasi hiperparameter modern yang mengimplementasikan pendekatan Bayesian menggunakan *Tree-structured Parzen Estimator (TPE)* dan telah terbukti efektif dalam meningkatkan performa model pada berbagai tugas pembelajaran mesin dan NLP [8].

Beberapa penelitian terkait menunjukkan potensi optimasi hiperparameter berbasis Bayesian dalam konteks NLP. Beberapa penelitian menunjukkan bahwa konfigurasi *fine-tuning* memiliki dampak signifikan terhadap kinerja dan stabilitas model BERT [2]. Pelatihan BERT sangat sensitif terhadap *random seed* dan konfigurasi hiperparameter, sehingga evaluasi tunggal tidak cukup merepresentasikan kualitas model [9]. Kajian lain menyebutkan mengenai model pralatih untuk NER menekankan pentingnya penyesuaian *fine-tuning* untuk mencapai performa optimal [10]. Sementara itu, Optuna diperkenalkan sebagai kerangka optimasi hiperparameter yang efisien dan fleksibel, yang kemudian diadopsi secara luas dalam penelitian pembelajaran mesin modern. Namun, sebagian besar penelitian tersebut tidak secara khusus mengevaluasi dampak optimasi hiperparameter terhadap stabilitas pelatihan model pada tugas NER Bahasa Indonesia.

Berdasarkan studi terdahulu, penelitian NER Bahasa Indonesia berbasis IndoBERT masih didominasi penggunaan konfigurasi fine-tuning bawaan atau adopsi parameter dari studi lain, sehingga belum tersedia bukti empiris yang secara sistematis memosisikan optimasi hiperparameter berbasis Optuna sebagai faktor utama peningkatan kinerja pada domain teks media sosial [2], [5], [6]. Di samping itu, pelaporan stabilitas pelatihan melalui evaluasi multi-seed dan analisis variansi kinerja masih terbatas, padahal aspek tersebut penting untuk memastikan reprodusibilitas dan keandalan temuan peningkatan performa [2], [9]. Penelitian ini berkontribusi dengan menyajikan evaluasi terukur penalaan hiperparameter IndoBERT menggunakan Optuna berbasis Tree-structured Parzen Estimator, mengaitkannya dengan metrik kinerja (precision, recall, dan F1-score) serta stabilitas pelatihan melalui pengujian multi-seed, sehingga menghasilkan rekomendasi konfigurasi fine-tuning yang lebih andal untuk pengembangan sistem NER Bahasa Indonesia pada teks media sosial [7], [10].

Oleh karena itu, penelitian ini mengevaluasi dampak optimasi hiperparameter berbasis Optuna terhadap kinerja IndoBERT pada tugas Named Entity Recognition Bahasa Indonesia menggunakan metrik precision, recall, dan F1-score. Stabilitas pelatihan dianalisis melalui evaluasi multi-seed dengan pelaporan nilai rerata dan standar deviasi performa. Penelitian ini diharapkan tidak hanya meningkatkan kinerja model, tetapi juga memberikan bukti empiris yang lebih reprodusibel mengenai sensitivitas hiperparameter pada IndoBERT, serta menghasilkan rekomendasi fine-tuning yang praktis dan andal untuk penerapan NER Bahasa Indonesia pada teks media sosial..

2. METODOLOGI PENELITIAN

2.1 Pendekatan Penelitian

Penelitian ini merupakan eksperimen kuantitatif yang mengevaluasi dampak optimasi hiperparameter berbasis Optuna terhadap kinerja dan stabilitas pelatihan model IndoBERT pada tugas Named Entity Recognition (NER) Bahasa Indonesia. Desain eksperimen menekankan keterulangan (reproducibility) dan validitas perbandingan melalui pengendalian variabel, serta pelaporan evaluasi yang lebih andal dengan pengujian multi-seed dan validasi silang [1], [2], [9].

2.2 Skema Penelitian

Penelitian ini mengevaluasi secara sistematis pengaruh optimasi hiperparameter berbasis Optuna terhadap kinerja dan stabilitas model IndoBERT pada tugas *Named Entity Recognition* (NER) Bahasa Indonesia. Metodologi penelitian disusun dalam beberapa tahapan yang saling terintegrasi, mulai dari pengumpulan data, prapemrosesan, pemodelan, optimasi, hingga evaluasi kinerja dan stabilitas model. Setiap tahapan dirancang untuk memastikan bahwa proses pelatihan dan pengujian model dilakukan secara objektif, terkontrol, dan dapat direproduksi.

Secara umum, alur metodologi penelitian ditunjukkan pada Gambar 1, yang menggambarkan rangkaian proses penelitian secara menyeluruh. Tahapan awal dimulai dengan pengumpulan dataset NER Bahasa Indonesia yang dianotasi menggunakan format CoNLL dan skema BIO. Data tersebut kemudian melalui tahap prapemrosesan yang mencakup pemetaan label, tokenisasi berbasis subword, penyelarasan label dengan subtoken, serta penyeragaman panjang urutan melalui proses *truncation* dan *padding*. Tahapan ini bertujuan untuk menyiapkan data agar kompatibel dengan arsitektur Transformer yang digunakan.

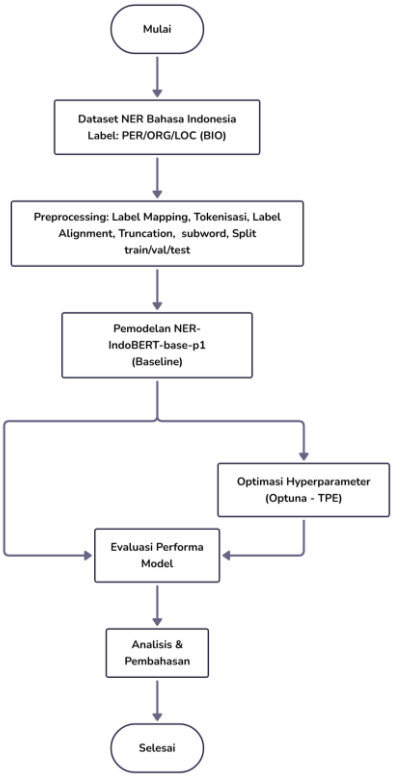
Selanjutnya, model IndoBERT diinisialisasi dan digunakan sebagai model dasar untuk tugas klasifikasi token. Penelitian ini menerapkan dua strategi eksperimen utama, yaitu pelatihan model dengan konfigurasi hiperparameter dasar

Copyright © 2026 The **Bambang Siswanto**, Page 222
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9545
Hal 221-233
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

(baseline) dan pelatihan dengan optimasi hiperparameter berbasis Optuna. Proses optimasi dilakukan secara iteratif melalui serangkaian *trial* yang mengevaluasi kinerja model pada data validasi, sehingga konfigurasi hiperparameter terbaik dapat diperoleh secara sistematis. Model terbaik hasil optimasi kemudian dievaluasi pada data uji untuk mengukur kinerja akhir.

Untuk memastikan keandalan hasil, evaluasi tidak hanya dilakukan berdasarkan satu kali pelatihan, tetapi juga diperkuat dengan analisis stabilitas pelatihan melalui pengujian multi-seed serta evaluasi robustness menggunakan *K-Fold Cross-Validation*. Hasil dari seluruh tahapan tersebut dianalisis secara komprehensif untuk menilai peningkatan kinerja, konsistensi performa, dan kemampuan generalisasi model. Dengan pendekatan metodologis ini, penelitian diharapkan mampu memberikan gambaran yang menyeluruh mengenai efektivitas optimasi hiperparameter Optuna dalam meningkatkan performa dan stabilitas model NER berbasis IndoBERT.



Gambar 1. Diagram Alir Penelitian

2.3 Dataset

Dataset pada penelitian ini bersumber dari Kaggle, yaitu "Bio Tagging NER Bahasa Indonesia (from Twitter/X)", yang berisi kumpulan teks berbahasa Indonesia hasil *crawl* dari platform Twitter/X dan telah dianotasi untuk tugas *Named Entity Recognition* (NER) [11]. Data disajikan mengikuti konvensi CoNLL, sehingga setiap token dipasangkan dengan label entitasnya, dan proses anotasi menerapkan skema BIO untuk membedakan token di luar entitas (O), token awal entitas (B-), serta token lanjutan di dalam entitas (I-). Fokus anotasi diarahkan pada entitas umum yang lazim digunakan dalam NER, yakni PERSON (PER), ORGANIZATION (ORG), dan LOCATION (LOC). Dataset kemudian digunakan dalam tiga pembagian data untuk eksperimen, yaitu data pelatihan sebanyak 8.153 kalimat, data validasi 1.019 kalimat untuk pemantauan kinerja selama pelatihan sekaligus penalaan hiperparameter, serta data uji 1.020 kalimat sebagai dasar evaluasi akhir pada data yang tidak terlihat selama pelatihan.

2.4 Prapemrosesan Data

Tahap prapemrosesan data dalam penelitian ini dirancang untuk memastikan bahwa data teks berbahasa Indonesia yang telah dianotasi dapat diproses secara optimal oleh model IndoBERT pada tugas *Named Entity Recognition* (NER). Mengingat arsitektur Transformer bekerja pada tingkat subtoken, setiap tahapan prapemrosesan dilakukan secara sistematis agar kesesuaian antara token dan label entitas tetap terjaga [12][13].

a. Pemetaan Label (Label Mapping)

Langkah awal prapemrosesan adalah melakukan pemetaan label entitas dari bentuk kategorikal ke representasi numerik. Dataset NER yang digunakan menerapkan skema BIO dengan label seperti *O*, *B-PER*, *I-PER*, *B-ORG*, *I-ORG*, *B-LOC*, dan *I-LOC*. Seluruh label tersebut dipetakan ke dalam bentuk indeks numerik melalui skema *label-to-id* (label2id), serta pemetaan balik *id-to-label* (id2label). Proses ini diperlukan agar label dapat diproses oleh model klasifikasi token dan memungkinkan interpretasi hasil prediksi model pada tahap evaluasi [14].

Copyright © 2026 The **Bambang Siswanto**, Page 223
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9545
Hal 221-233
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

b. Tokenisasi Berbasis Subword

Setelah pemetaan label, setiap kalimat dalam dataset dilakukan proses tokenisasi menggunakan *tokenizer* bawaan IndoBERT yang berbasis *WordPiece*. Tokenisasi berbasis subword memungkinkan model menangani variasi morfologi dan kosakata Bahasa Indonesia secara lebih fleksibel, termasuk kata tidak baku atau kata dengan imbuhan. Pada tahap ini, satu token kata dapat dipecah menjadi beberapa subtoken, yang selanjutnya diproses sebagai unit input oleh model Transformer[15].

c. Penyelarasan Label dengan Subtoken (Label Alignment)

Tokenisasi berbasis subword menimbulkan tantangan pada tugas NER karena anotasi awal diberikan pada tingkat kata. Oleh karena itu, penelitian ini menerapkan prosedur *tokenize and align labels* untuk menyelaraskan label entitas dengan subtoken hasil tokenisasi. Label entitas asli hanya diberikan pada subtoken pertama dari setiap kata, sedangkan subtoken lanjutan diberi label khusus yang diabaikan dalam perhitungan fungsi loss. Pendekatan ini bertujuan untuk mencegah bias pelatihan akibat pembelahan subtoken dan memastikan bahwa pembelajaran model tetap merepresentasikan anotasi entitas pada tingkat kata secara akurat [16].

d. Pemotongan Panjang Urutan (Truncation)

Untuk menjaga efisiensi komputasi dan keseragaman dimensi input, setiap urutan token dipotong hingga panjang maksimum 128 token. Batas maksimum ini dipilih sebagai kompromi antara kebutuhan konteks kalimat dan keterbatasan sumber daya komputasi, mengingat sebagian besar kalimat dalam dataset memiliki panjang yang relatif pendek. Pemotongan ini memastikan bahwa input yang melebihi panjang maksimum tidak menyebabkan peningkatan kompleksitas komputasi secara signifikan [14].

e. Padding dan Penyusunan Batch Data

Apabila panjang urutan token lebih pendek dari batas maksimum, dilakukan proses *padding* untuk menyeragamkan panjang input. Proses ini ditangani menggunakan *data collator* khusus untuk tugas *token classification*, sehingga token *padding* tidak memengaruhi perhitungan fungsi loss maupun evaluasi model. Dengan mekanisme ini, setiap batch data memiliki dimensi yang seragam dan dapat diproses secara efisien oleh model IndoBERT selama pelatihan dan evaluasi [17].

f. Pembagian Dataset untuk Pelatihan dan Evaluasi

Dataset yang telah melalui tahap tokenisasi dan penyelarasan label kemudian disusun ke dalam subset pelatihan, validasi, dan pengujian. Data pelatihan digunakan untuk proses pembaruan parameter model, data validasi dimanfaatkan untuk memantau kinerja model serta sebagai dasar optimasi hiperparameter, sedangkan data pengujian digunakan untuk mengevaluasi performa akhir model pada data yang tidak pernah dilihat selama proses pelatihan. *Subset pelatihan–validasi–uji digunakan untuk menjaga objektivitas evaluasi dan mengurangi risiko data leakage, dengan validasi dipakai untuk pemantauan kinerja dan dasar optimasi hiperparameter [1], [2].*

2.5 IndoBert

Penelitian ini menggunakan model IndoBERT dengan varian *indobenchmark/indobert-base-p1*, yaitu model BERT berarsitektur *Base* yang telah melalui tahap *pre-training* pada korpus besar Bahasa Indonesia dan dikembangkan dalam kerangka kerja IndoNLP/IndoNLU [18]. Model ini dirancang untuk menangkap karakteristik linguistik Bahasa Indonesia secara lebih representatif dibandingkan model BERT multibahasa, khususnya pada tugas pemahaman teks tingkat token seperti *Named Entity Recognition* (NER) [19]. Pemilihan varian *Base* didasarkan pada pertimbangan keseimbangan antara kapasitas representasi kontekstual dan efisiensi komputasi, sehingga eksperimen dapat dijalankan secara stabil pada lingkungan dengan sumber daya terbatas, seperti Google Colab dengan GPU kelas menengah.

IndoBERT mengadopsi arsitektur Transformer dengan mekanisme *bidirectional self-attention*, yang memungkinkan model memanfaatkan konteks token secara dua arah. Kemampuan ini menjadi krusial dalam tugas NER, karena penentuan label entitas pada suatu token sering kali dipengaruhi oleh kata sebelum dan sesudahnya. Dalam penelitian ini, IndoBERT digunakan sebagai model dasar (baseline) dengan menambahkan lapisan *token classification* di atas encoder Transformer untuk memetakan representasi kontekstual token ke dalam label BIO yang mencakup entitas PERSON (PER), ORGANIZATION (ORG), dan LOCATION (LOC).

Teks masukan diproses menggunakan *tokenizer* WordPiece bawaan IndoBERT, yang bekerja pada tingkat subtoken untuk menangani variasi morfologi dan kosakata Bahasa Indonesia. Panjang maksimum urutan token ditetapkan sebesar 128 token, sesuai dengan karakteristik dataset dan keterbatasan memori GPU. Strategi ini memastikan bahwa konteks semantik utama tetap terjaga tanpa meningkatkan kompleksitas komputasi secara berlebihan. Proses *truncation* dan *padding* dilakukan secara otomatis agar seluruh input memiliki dimensi yang seragam.

Proses *fine-tuning* IndoBERT diimplementasikan menggunakan kerangka kerja Hugging Face Transformers dengan memanfaatkan kelas AutoModelForTokenClassification dan Trainer [2]. Seluruh parameter IndoBERT dan lapisan klasifikasi diperbarui secara end-to-end menggunakan fungsi *cross-entropy loss* pada level token, dengan mengabaikan token *padding* dan subtoken lanjutan yang tidak memiliki label entitas asli. Pendekatan ini memastikan bahwa pembaruan parameter hanya dipengaruhi oleh token yang relevan terhadap tugas NER.

Tabel 1 adalah konfigurasi parameter untuk pelatihan baseline IndoBERT [6]. Hasil evaluasi baseline menunjukkan bahwa IndoBERT mampu mencapai kinerja yang kompetitif pada tugas NER Bahasa Indonesia, namun masih menunjukkan keterbatasan dalam hal performa optimal dan stabilitas pelatihan. Temuan ini menjadi dasar empiris

This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9545
Hal 221-233
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

untuk menerapkan optimasi hiperparameter berbasis Optuna pada tahap selanjutnya, guna meningkatkan kinerja dan konsistensi hasil pelatihan. Dengan demikian, Subbab ini berperan sebagai jembatan metodologis yang menghubungkan pemilihan dan konfigurasi IndoBERT dengan analisis hasil eksperimen baseline dan optimasi yang disajikan pada bagian Hasil dan Pembahasan.

Tabel 1. Parameter Model Baseline

Parameter	Nilai
Model Pra-latih	indobenchmark/indobert-base-p1
Varian Model	Base
Optimizer	AdamW
Jumlah Epoch	3
Batch Size (Training)	8
Batch Size (Evaluasi)	8
Learning Rate	2×10^{-5}
Scheduler	Linear warm-up
Panjang Maksimal Token	128
Logging Steps	50
Perangkat	Google Colab (GPU T4)
Framework	Hugging Face Transformers (Trainer API)

2.6 Hiperparameter Optuna

Optimasi hiperparameter merupakan komponen penting dalam penelitian ini karena performa *fine-tuning* model IndoBERT pada tugas *Named Entity Recognition* (NER) sangat dipengaruhi oleh konfigurasi pelatihan. Pemilihan nilai hiperparameter yang kurang tepat dapat menyebabkan konvergensi yang lambat, ketidakstabilan proses pelatihan, serta fluktuasi kinerja antar percobaan. Sejumlah penelitian menunjukkan bahwa pendekatan penalaan manual atau penggunaan konfigurasi bawaan sering kali belum mampu menghasilkan performa optimal pada model berbasis Transformer, khususnya pada tugas NLP dengan kompleksitas tinggi seperti NER [20]. Oleh karena itu, penelitian ini menerapkan Optuna sebagai kerangka *hyperparameter optimization* yang mengadopsi pendekatan *Bayesian Optimization* berbasis Tree-structured Parzen Estimator (TPE). Pendekatan ini memungkinkan proses pencarian hiperparameter dilakukan secara adaptif dengan memanfaatkan informasi dari hasil percobaan sebelumnya, sehingga lebih efisien dibandingkan pencarian acak (*random search*) maupun *grid search* konvensional [7].

a. Formulasi Masalah Optimasi

Proses optimasi diformulasikan sebagai masalah pencarian konfigurasi hiperparameter θ yang memaksimalkan kinerja model pada data validasi. Dalam penelitian ini, fungsi objektif didefinisikan sebagai F1-score validasi, karena metrik tersebut paling representatif untuk tugas NER yang umumnya memiliki distribusi label tidak seimbang [8]. Secara konseptual, Optuna melakukan iterasi pada sejumlah *trial* untuk mencari konfigurasi yang menghasilkan nilai F1-score terbaik.

Pada setiap *trial*, model IndoBERT dilatih menggunakan konfigurasi hiperparameter yang diusulkan Optuna dan kemudian dievaluasi pada data validasi. Nilai F1-score hasil evaluasi tersebut digunakan sebagai umpan balik untuk memperbarui strategi pencarian pada *trial* berikutnya, sehingga proses optimasi berlangsung secara iteratif dan terarah .

b. Ruang Pencarian Hiperparameter (Search Space)

Ruang pencarian hiperparameter dalam penelitian ini mencakup parameter-parameter yang secara empiris diketahui berpengaruh signifikan terhadap *fine-tuning* model Transformer [21]. Hiperparameter yang dioptimasi mengikuti konfigurasi pada notebook penelitian, meliputi *learning rate*, *batch size*, jumlah *epoch*, *weight decay*, *warmup ratio*, *max gradient norm*, dan *label smoothing factor*. *Learning rate* dicari pada rentang logaritmik untuk menangkap sensitivitas model terhadap skala pembaruan bobot, sedangkan *weight decay* dan *label smoothing* berfungsi sebagai mekanisme regularisasi untuk mengurangi overfitting. Sementara itu, *warmup ratio* dan *max gradient norm* berperan penting dalam menjaga stabilitas gradien selama proses pelatihan. Pemilihan ruang pencarian ini bertujuan untuk menyeimbangkan aspek konvergensi, regularisasi, dan stabilitas pelatihan, mengingat tugas NER memerlukan prediksi token-level yang sangat sensitif terhadap konteks.

c. Mekanisme Optimasi Optuna (TPE Sampler)

Optuna menggunakan TPE sampler yang memodelkan distribusi probabilitas dari konfigurasi hiperparameter dengan performa tinggi dan rendah berdasarkan nilai objektif yang diperoleh pada *trial* sebelumnya. Alih-alih mencoba kombinasi hiperparameter secara acak, TPE membangun dua distribusi probabilistik, yaitu $l(\theta)$ untuk konfigurasi dengan performa baik dan $g(\theta)$ untuk konfigurasi dengan performa kurang baik. Konfigurasi hiperparameter baru kemudian dipilih dengan memaksimalkan rasio $l(\theta)/g(\theta)$, sehingga pencarian difokuskan pada wilayah ruang parameter yang lebih menjanjikan. Pendekatan ini telah terbukti mampu meningkatkan efisiensi pencarian hiperparameter pada berbagai tugas pembelajaran mendalam, termasuk NLP berbasis Transformer.

d. Integrasi Optuna dengan Proses Fine-Tuning IndoBERT

This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9545
Hal 221-233
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Dalam implementasinya, setiap *trial* Optuna menjalankan prosedur *fine-tuning* IndoBERT yang identik, namun dengan konfigurasi hiperparameter yang berbeda. Model dibangun menggunakan AutoModelForTokenClassification, sedangkan proses pelatihan dikelola melalui kelas Trainer pada pustaka Hugging Face Transformers [6][22]. Selama pelatihan, performa model dievaluasi pada data validasi di setiap epoch, dan model terbaik dipilih berdasarkan nilai F1-score tertinggi. Untuk mencegah pelatihan yang tidak efisien dan menghindari overfitting, penelitian ini juga menerapkan mekanisme early stopping, sehingga pelatihan dapat dihentikan lebih awal apabila kinerja validasi tidak mengalami peningkatan dalam beberapa epoch berturut-turut.

e. Kriteria Pemilihan Model Terbaik dan Evaluasi Lanjutan

Hasil optimasi Optuna menghasilkan konfigurasi hiperparameter terbaik (*Optuna-best*) berdasarkan nilai F1-score validasi tertinggi. Konfigurasi ini kemudian digunakan untuk melatih ulang model dan dievaluasi pada data uji untuk memperoleh performa akhir. Selain evaluasi satu kali pelatihan, penelitian ini juga menerapkan evaluasi multi-seed serta K-Fold Cross-Validation untuk menilai stabilitas dan robustness model. Pendekatan evaluasi lanjutan ini bertujuan memastikan bahwa konfigurasi hiperparameter yang dipilih tidak hanya unggul secara kinerja, tetapi juga konsisten dan dapat direproduksi pada berbagai kondisi pelatihan [2].

2.7 Desain Eksperimen dan Skenario Pengujian

Desain eksperimen dalam penelitian ini disusun untuk mengevaluasi secara objektif dampak optimasi hiperparameter berbasis Optuna terhadap kinerja dan stabilitas model IndoBERT pada tugas *Named Entity Recognition* (NER) Bahasa Indonesia. Pendekatan eksperimental ini sejalan dengan praktik evaluasi model *deep learning* modern yang menekankan perbandingan sistematis antara konfigurasi dasar (*baseline*) dan model teroptimasi [1]. Untuk tujuan tersebut, penelitian ini membandingkan dua skenario utama, yaitu model baseline IndoBERT dan model IndoBERT hasil optimasi Optuna, dengan seluruh variabel lain dikendalikan agar perbandingan yang dihasilkan bersifat adil (*fair comparison*).

Pada skenario baseline, model IndoBERT dilatih menggunakan konfigurasi hiperparameter standar yang umum digunakan dalam *fine-tuning* model Transformer, tanpa proses optimasi otomatis. Pendekatan ini lazim digunakan sebagai titik acuan dalam penelitian NLP untuk mengukur peningkatan kinerja yang dihasilkan oleh metode optimasi lanjutan [23]. Seluruh proses pelatihan baseline menggunakan arsitektur model, tokenizer, dataset, dan pembagian data yang sama dengan skenario optimasi, sehingga perbedaan kinerja yang diperoleh dapat diatribusikan secara langsung pada strategi optimasi hiperparameter.

Pada skenario optimasi, Optuna digunakan untuk mencari kombinasi hiperparameter yang optimal melalui serangkaian *trial* berbasis *Bayesian Optimization* dengan *Tree-structured Parzen Estimator (TPE)* sampler. Pendekatan ini telah banyak digunakan dalam penelitian pembelajaran mendalam karena kemampuannya mengeksplorasi ruang parameter secara efisien dengan memanfaatkan hasil evaluasi sebelumnya. Setiap *trial* merepresentasikan satu konfigurasi hiperparameter yang digunakan untuk melatih model IndoBERT dan dievaluasi pada data validasi menggunakan metrik F1-score. Konfigurasi dengan nilai F1-score validasi tertinggi dipilih sebagai Optuna-best, yang kemudian digunakan pada tahap evaluasi akhir.

Untuk menjaga konsistensi eksperimen dan memastikan reprodusibilitas, beberapa parameter ditetapkan tetap (*fixed settings*) pada seluruh skenario, meliputi panjang maksimum urutan token, skema tokenisasi dan pelabelan BIO, pembagian dataset (train–validation–test), serta metrik evaluasi utama. Pengendalian variabel semacam ini merupakan praktik yang direkomendasikan dalam desain eksperimen pembelajaran mesin agar kesimpulan yang diperoleh bersifat valid dan dapat dibandingkan dengan penelitian sebelumnya.

2.8 Evaluasi Model

Evaluasi model dalam penelitian ini dilakukan secara bertahap dan komprehensif untuk mengukur kinerja, stabilitas, dan robustness model IndoBERT pada tugas NER Bahasa Indonesia. Mengingat NER merupakan tugas klasifikasi token dengan distribusi label yang tidak seimbang, pemilihan metrik evaluasi dan prosedur pengujian mengikuti rekomendasi penelitian NLP mutakhir yang menekankan penggunaan metrik berbasis *precision*, *recall*, dan *F1-score* dibandingkan akurasi semata [24].

a. Evaluasi Kinerja Utama

Penelitian ini mengevaluasi performa model menggunakan tiga metrik utama, yaitu Precision, Recall, dan F1-score, yang dihitung berdasarkan hasil klasifikasi token dalam tugas NER [14]. Precision mengukur ketepatan model dalam memprediksi entitas yang benar dari semua prediksi yang dihasilkan, sedangkan recall mengukur kemampuan model dalam menemukan semua entitas yang relevan dari data sebenarnya. F1-score kemudian dihitung sebagai rata-rata harmonik dari precision dan recall, sehingga memberikan gambaran seimbang atas kedua metrik tersebut. Persamaan presisi, recall, dan F1-score sebagaimana persamaan berikut :

$$Precision = \frac{TP}{TP+FP}\ x\ 100\% \quad (1)$$

$$Recall = \frac{TP}{TP+FN}\ x\ 100\% \quad (2)$$

$$F1\ Score = 2x\frac{Precision\ x\ Recall}{Precision+\ Recall} \quad (3)$$

Copyright © 2026 The **Bambang Siswanto**, Page 226
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9545
Hal 221-233
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Selain evaluasi global, analisis kinerja juga dilakukan pada tingkat entitas untuk kelas PERSON (PER), ORGANIZATION (ORG), dan LOCATION (LOC). Evaluasi per entitas ini penting untuk mengidentifikasi variasi performa model terhadap jenis entitas yang berbeda, sebagaimana banyak dilakukan dalam studi NER berbasis BERT. Hasil evaluasi ini disajikan dan dibahas secara rinci pada pembahasan.

b. Evaluasi Stabilitas Pelatihan (Multi-Seed Evaluation)

Untuk menilai konsistensi performa model terhadap variasi inisialisasi acak, penelitian ini menerapkan evaluasi multi-seed, yaitu dengan mengulangi proses pelatihan dan evaluasi menggunakan beberapa nilai *random seed*. Pendekatan ini direkomendasikan dalam literatur *fine-tuning* Transformer karena performa model diketahui sensitif terhadap inisialisasi bobot dan urutan data pelatihan.

Nilai F1-score dari setiap percobaan dianalisis untuk memperoleh nilai rata-rata dan standar deviasi, yang digunakan sebagai indikator stabilitas pelatihan. Standar deviasi yang lebih kecil menunjukkan bahwa model menghasilkan performa yang lebih konsisten dan dapat direproduksi. Hasil evaluasi multi-seed disajikan dalam bentuk distribusi F1-score dan *boxplot*, sebagaimana disarankan dalam penelitian evaluasi stabilitas model *deep learning* [13].

c. Evaluasi Stabilitas Pelatihan (Multi-Seed Evaluation)

Sebagai evaluasi lanjutan, penelitian ini menerapkan *K-Fold Cross-Validation* untuk menilai robustness dan kemampuan generalisasi model terhadap variasi pembagian data. Teknik ini banyak digunakan dalam penelitian pembelajaran mesin untuk mengurangi bias yang timbul akibat satu pembagian data tertentu. Pada setiap iterasi, satu fold digunakan sebagai data validasi sementara fold lainnya digunakan sebagai data pelatihan, hingga seluruh fold berperan sebagai data validasi.

Nilai metrik evaluasi dari setiap fold kemudian dirata-ratakan untuk memperoleh estimasi kinerja yang lebih stabil. Pendekatan ini memberikan bukti tambahan bahwa konfigurasi hiperparameter terbaik hasil Optuna memiliki performa yang konsisten dan tidak bergantung pada satu skenario pengujian tertentu, sehingga memperkuat klaim generalisasi model pada tugas NER Bahasa Indonesia.

3. HASIL DAN PEMBAHASAN

Penelitian ini menyajikan hasil eksperimen serta pembahasan komprehensif mengenai kinerja model IndoBERT pada tugas *Named Entity Recognition* (NER) Bahasa Indonesia, baik sebelum maupun sesudah dilakukan optimasi hiperparameter berbasis Optuna. Penyajian hasil disusun secara sistematis untuk mencerminkan tahapan metodologi yang telah dijelaskan pada Bab 2, sehingga hubungan antara desain eksperimen, prosedur evaluasi, dan temuan empiris dapat ditelusuri secara jelas.

3.1 Hasil Penelitian

3.1.1 Prapemrosesan Data

Prapemrosesan data dilakukan untuk memastikan dataset Twitter/X berformat CoNLL dengan skema BIO dapat digunakan secara konsisten pada tugas *token classification* berbasis transformer. Tahap awal mencakup validasi struktur data (token–label dan batas kalimat), pemetaan label BIO (O, B-*,* I-) ke indeks numerik (label2id/id2label), serta pemeriksaan konsistensi anotasi agar ketidaksesuaian label (misalnya I-* tanpa didahului B-* pada entitas yang sama) tidak memengaruhi proses pelatihan dan evaluasi. Langkah ini krusial karena kualitas pelabelan BIO berpengaruh langsung terhadap ketepatan *boundary* dan konsistensi span entitas [1], [10], [11].

Selanjutnya, teks ditokenisasi menggunakan tokenizer IndoBERT berbasis subword, kemudian dilakukan penyelarasan label pada tingkat subtoken agar anotasi pada level kata tetap terwakili setelah pemecahan token [20], [22]. Fitur input dibentuk menjadi *input_ids* dan *attention_mask* dengan mekanisme padding/truncation pada panjang sekuens tertentu; label untuk token padding diabaikan dalam perhitungan *loss* agar tidak menimbulkan bias. Dengan pipeline ini, data siap digunakan untuk fine-tuning baseline dan Optuna secara adil, sedangkan keandalan temuan diperkuat pada tahap evaluasi melalui multi-seed dan K-Fold Cross-Validation [2], [9], [21].

3.1.2 Model IndoBERT Baseline

Setelah tahap pelatihan baseline, penelitian dilanjutkan dengan optimasi hiperparameter menggunakan Optuna dengan sampler Tree-structured Parzen Estimator (TPE), yaitu pendekatan optimasi berbasis model yang menyesuaikan strategi pencarian secara adaptif berdasarkan performa trial sebelumnya [7]. Pada setiap trial, Optuna mengevaluasi satu konfigurasi hiperparameter dan menggunakan F1-score pada data validasi sebagai *objective function*, sehingga pencarian difokuskan pada kombinasi parameter yang paling berpotensi meningkatkan kemampuan generalisasi model.

Berdasarkan hasil optimasi, konfigurasi terbaik (Optuna) menghasilkan performa evaluasi pada data uji sebesar Precision = 0,9338, Recall = 0,9312, F1-score = 0,9325, dan Accuracy = 0,9854 (Tabel 3). Konfigurasi Optuna yang terpilih adalah learning rate = $5{,}7745\times10^{-5}$, batch size = 8, epochs = 5, weight decay = 0,05, warmup ratio = 0,1, max_grad_norm = 0,1625, dan label smoothing factor = 0,1, dengan n-trial = 50.

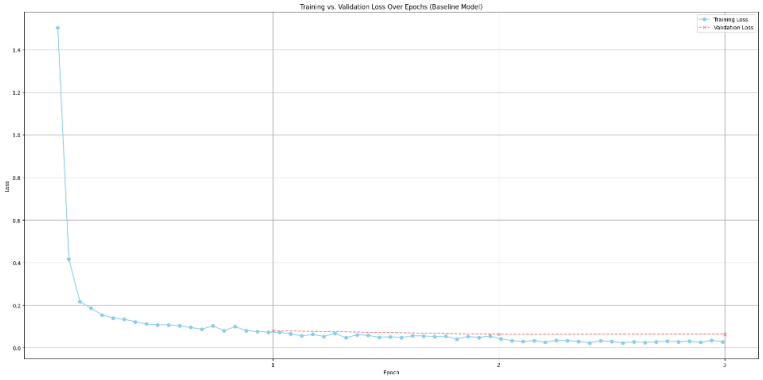
Copyright © 2026 The **Bambang Siswanto**, Page 227
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9545
Hal 221-233
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tabel 2. Evaluation Metric Model Baseline

Model IndoBERT	Metric Evaluasi			
	Precision	Recall	F1 Score	Accuracy
Baseline	0.921099	0.929453	0.925257	0.983744

Dari sisi dinamika pelatihan, Gambar 2 memperlihatkan penurunan training loss yang tajam pada fase awal dan kemudian melandai, sementara validation loss relatif stabil serta mengikuti tren penurunan yang serupa. Pola ini menunjukkan proses fine-tuning berjalan konvergen dan memasuki kondisi plateau pada epoch berikutnya [3].



Gambar 2. Kurva Training Loss dan Validation Loss Baseline

3.1.3 Optimasi Hiperparameter Menggunakan Optuna

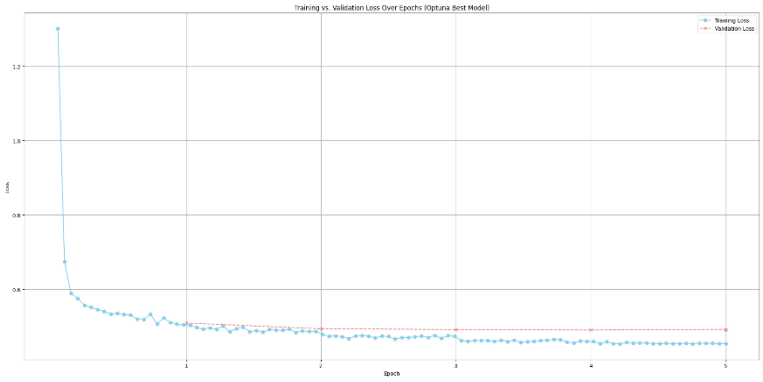
Setelah tahap pelatihan baseline, penelitian dilanjutkan dengan optimasi hiperparameter menggunakan Optuna dengan sampler Tree-structured Parzen Estimator (TPE), yaitu pendekatan optimasi berbasis model yang menyesuaikan strategi pencarian secara adaptif berdasarkan performa trial sebelumnya [7]. Pada setiap trial, Optuna mengevaluasi satu konfigurasi hiperparameter dan menggunakan F1-score pada data validasi sebagai objective function, sehingga pencarian difokuskan pada kombinasi parameter yang paling berpotensi meningkatkan kemampuan generalisasi model.

Berdasarkan hasil optimasi, konfigurasi terbaik (Optuna) menghasilkan performa evaluasi pada data uji sebesar Precision = 0,9338, Recall = 0,9312, F1-score = 0,9325, dan Accuracy = 0,9854 sebagaimana ditunjukkan Tabel 3. Konfigurasi Optuna yang terpilih adalah learning rate = $5{,}7745 \times 10^{-5}$, batch size = 8, epochs = 5, weight decay = 0,05, warmup ratio = 0,1, max_grad_norm = 0,1625, dan label smoothing factor = 0,1, dengan n-trial = 50.

Tabel 3. Evaluation Metric Model Optuna

Model IndoBERT	Metric Evaluasi			
	Precision	Recall	F1 Score	Accuracy
Optuna	0.933805	0.931217	0.932509	0.985449

Dari sisi dinamika pelatihan, Gambar 3 memperlihatkan penurunan training loss yang tajam pada epoch awal kemudian melandai dan stabil pada epoch berikutnya. Sementara itu, validation loss berada pada kisaran yang relatif datar dan tidak menunjukkan tren meningkat hingga akhir pelatihan, yang mengindikasikan tidak terjadi overfitting yang signifikan serta model mempertahankan kemampuan generalisasi yang baik [1].

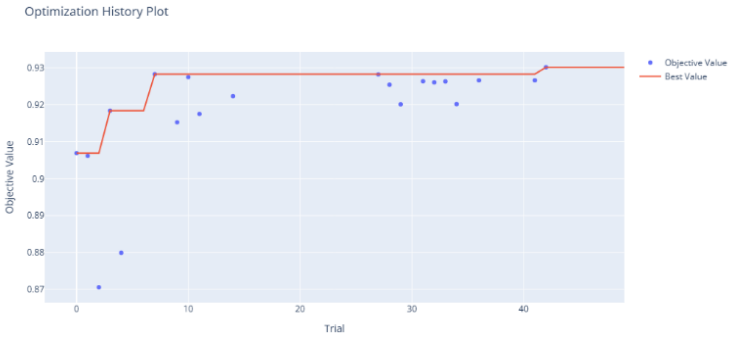


Gambar 3. Kurva Training Loss dan Validation Loss Model dengan Optuna

Untuk memantau dinamika optimasi, visualisasi optimization history plot pada Gambar 4 menunjukkan peningkatan nilai *best objective* (F1 validasi) secara bertahap seiring bertambahnya trial, kemudian memasuki pola plateau setelah kandidat terbaik tercapai.

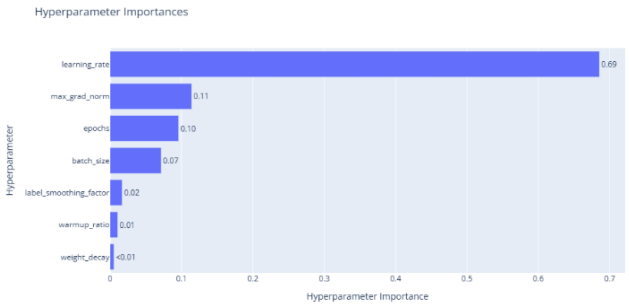
Copyright © 2026 The **Bambang Siswanto**, Page 228
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9545
Hal 221-233
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



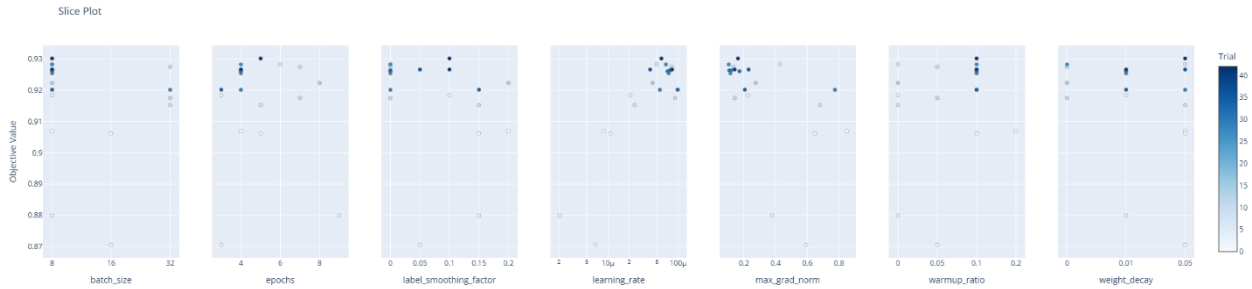
Gambar 4. Optimization history plot Optuna

Analisis hyperparameter importances pada Gambar 5 menunjukkan bahwa learning rate merupakan faktor paling dominan terhadap F1-score validasi, sedangkan *max_grad_norm, epochs, dan batch size* berperan sebagai pengendali stabilitas optimisasi dan durasi pembelajaran.



Gambar 5. Grafik *Hyperparameter importance Optuna*

Selanjutnya, slice plot pada Gambar 6 memperlihatkan bahwa trial dengan F1 tinggi terkonsentrasi pada rentang learning rate moderat; nilai learning rate yang terlalu kecil maupun terlalu besar cenderung berkorelasi dengan performa yang lebih rendah..



Gambar 6. Grafik *Slice Plot* Optuna

3.1.4 Perbandingan Kinerja Per Entitas Baseline dan Optuna

Perbandingan kinerja antara model baseline dan Optuna menunjukkan adanya peningkatan pada metrik evaluasi per label. Tabel 4 menyajikan metrik precision, recall, dan F1-score untuk setiap label BIO. Hasil menunjukkan bahwa label **O** mendominasi jumlah token (support = 28.941), sedangkan label entitas (B-/*I*-) memiliki support yang lebih kecil, sehingga pembacaan performa perlu mempertimbangkan metrik per label dan tidak hanya metrik agregat [1], [10]. Pada label **O**, kedua konfigurasi memperoleh nilai sangat tinggi (Baseline F1 = 0,9918; Optuna F1 = 0,9925) dengan pergeseran kecil pada recall O yang lebih tinggi pada Optuna.

Tabel 4. Evaluation Metric Per Entitas

Label	Baseline			Optuna		
	Precision	Recall	F1	Precision	Recall	F1
O	0.9923	0.9913	0.9918	0.9919	0.9931	0.9925
B-LOC	0.9587	0.9617	0.9602	0.9648	0.9663	0.9656
B-ORG	0.9403	0.9397	0.94	0.9488	0.939	0.9439
B-PER	0.9286	0.919	0.9237	0.9351	0.919	0.927

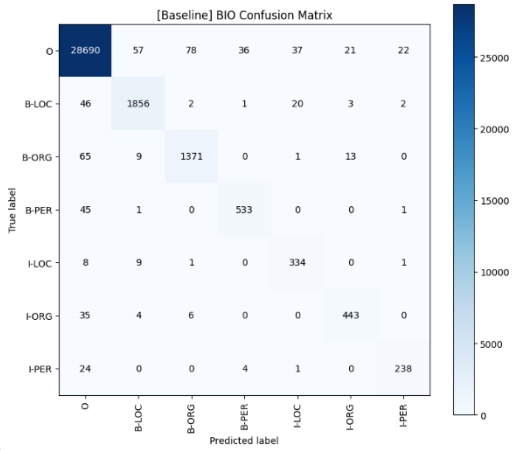
Copyright © 2026 The **Bambang Siswanto**, Page 229
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9545
Hal 221-233
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

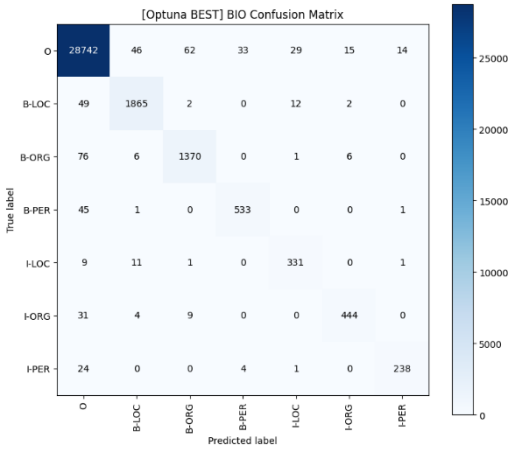
I-LOC	0.8499	0.9462	0.8954	0.885	0.9377	0.9106
I-ORG	0.9229	0.9078	0.9153	0.9507	0.9098	0.9298
I-PER	0.9015	0.8914	0.8964	0.937	0.8914	0.9136

Pada label lanjutan entitas **I-LOC**, **I-ORG**, dan **I-PER**, peningkatan lebih menonjol. Untuk **I-LOC** (support = 353), baseline menunjukkan ketidakseimbangan precision–recall (precision = 0,8499; recall = 0,9462), sedangkan Optuna meningkatkan precision menjadi 0,8850 dengan recall tetap tinggi (0,9377), sehingga F1 meningkat dari 0,8954 menjadi 0,9106. Pola serupa terlihat pada **I-ORG** (F1 meningkat dari 0,9153 menjadi 0,9298) dan **I-PER** (F1 meningkat dari 0,8964 menjadi 0,9136).

Bukti kuantitatif pada Tabel 4 diperkuat oleh confusion matrix pada **Gambar 7** (baseline) dan **Gambar 8** (Optuna). Kedua gambar menunjukkan bahwa kesalahan dominan lebih sering terjadi pada token yang berada di sekitar batas entitas, termasuk token entitas yang diprediksi sebagai **O** dan pergeseran label antara **B-** dan **I-**.



Gambar 7. Confusion Matrix IndoBERT Baseline



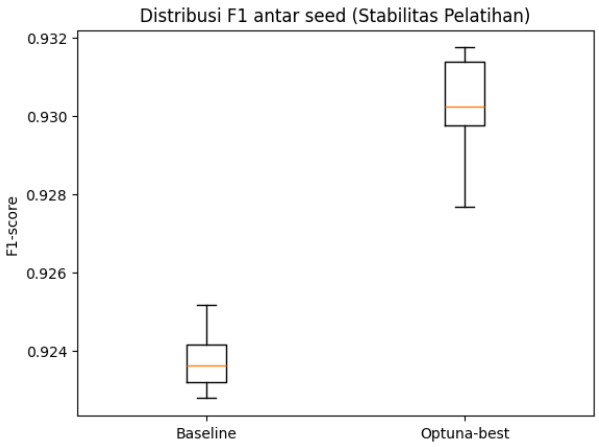
Gambar 8. Confusion Matrix Indobert Optuna

3.1.5 Stabilitas Pelatihan (Multi-Seed Analysis)

Untuk menilai konsistensi performa model, evaluasi dilakukan menggunakan beberapa nilai *random seed*. Secara kuantitatif, Tabel 5 menunjukkan bahwa model baseline memperoleh precision 0,9213 ± 0,0012, recall 0,9263 ± 0,0015, dan F1-score 0,9238 ± 0,0009, sedangkan Optuna mencapai precision 0,9287 ± 0,0032, recall 0,9317 ± 0,0009, dan F1-score 0,9302 ± 0,0016. Visualisasi boxplot pada Gambar 9 memperlihatkan sebaran F1-score Optuna yang lebih rapat dibandingkan baseline, sehingga variasi kinerja antar-seed terlihat lebih terkendali.

Tabel 5. Multi-Seed Analysis

Metode	Precision (mean±std)	Recall (mean±std)	F1 (mean±std)
Baseline	0.9213 ± 0.0012	0.9263 ± 0.0015	0.9238 ± 0.0009
Optuna	0.9287 ± 0.0032	0.9317 ± 0.0009	0.9302 ± 0.0016



Gambar 9. Boxplot distribusi F1

Berdasarkan uji signifikansi berpasangan pada skor F1 multi-seed yang dirangkum pada Tabel 6, perbandingan baseline vs Optuna menghasilkan paired t-test dengan *t-statistic* = −10,0759 dan *p-value* = 0,0005. Sebagai pembanding non-parametrik, Wilcoxon signed-rank test menghasilkan *statistic* = 0,0000 dan *p-value* = 0,0625.

Copyright © 2026 The **Bambang Siswanto**, Page 230
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9545
Hal 221-233
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tabel 6. Uji signifikansi model

Metode	Paired T-Test		Wilcoxon	
	T-statistic	p-value	Statistic	p-value
Baseline vs Optuna Best	-10.0759	0.0005	0.0000	0.0625

3.1.6 Evaluasi Robustness Menggunakan K-Fold Cross-Validation

Evaluasi K-Fold Cross-Validation (K-Fold CV) dilakukan untuk menilai robustness dan kemampuan generalisasi model pada variasi pembagian data, sehingga mengurangi ketergantungan pada satu skema train–test split tertentu. Hasil pengujian pada lima fold untuk konfigurasi Optuna dirangkum pada Tabel 7. Secara keseluruhan, performa model menunjukkan nilai tinggi dan relatif stabil pada seluruh fold, dengan precision berada pada rentang 0,9716–0,9788, recall 0,9778–0,9813, F1-score 0,9753–0,9801, dan accuracy 0,9949–0,9960. Rata-rata K-Fold yang diperoleh adalah precision = 0,9758, recall = 0,9796, F1 = 0,9777, dan accuracy = 0,9954.

Tabel 7. Hasil K-Fold Cross-Validation (Optuna)

Fold	Precision	Recall	F1	Accuracy
Fold 1	0.9751	0.9793	0.9772	0.9952
Fold 2	0.9788	0.9813	0.9801	0.996
Fold 3	0.9716	0.979	0.9753	0.9949
Fold 4	0.9765	0.9778	0.9771	0.9951
Fold 5	0.9768	0.9803	0.9786	0.9957
Average (K-Fold)	0.9758	0.9796	0.9777	0.9954

3.2 Pembahasan

3.2.1 Kinerja Baseline IndoBERT

Hasil baseline sebagaimana tabel 2 menunjukkan bahwa IndoBERT mampu memberikan performa awal yang kompetitif pada NER berbahasa Indonesia, sejalan dengan temuan bahwa model pra-latih monolingual Indonesia cenderung kuat pada berbagai tugas NLP Indonesia [5], [19]. Meski demikian, selisih kecil antara precision dan recall pada baseline mengindikasikan adanya trade-off yang masih dapat ditingkatkan, terutama pada konteks NER yang menuntut ketepatan *boundary* serta konsistensi span pada skema BIO/IOB [1], [10], [24]. Indikasi plateau pada kurva training–validation loss baseline seperti ditunjukkan pada gambar 2 juga menguatkan bahwa konfigurasi default mulai mencapai batas performa pada ruang hiperparameter yang digunakan, sehingga penalaan yang lebih sistematis berpeluang mendorong peningkatan lebih lanjut [2], [9].

3.2.2 Optimasi Hiperparameter Optuna

Peningkatan kinerja setelah optimasi (Tabel 3) memperkuat bahwa fine-tuning transformer sensitif terhadap konfigurasi optimisasi, terutama *learning rate*, regularisasi, dan mekanisme stabilisasi pembelajaran [2], [9]. Stabilitas konvergensi pada Optuna ditunjukkan oleh kurva training–validation loss yang relatif terkendali tanpa indikasi kenaikan validation loss (Gambar 3), yang mengindikasikan kemampuan generalisasi yang lebih baik dibanding baseline (Gambar 2). Dari sisi proses, pola konvergensi pada optimization history (Gambar 4) memperlihatkan peningkatan *best objective value* yang bertahap hingga memasuki fase plateau, yang konsisten dengan karakteristik optimasi berbasis model seperti TPE [7].

Temuan ini diperkuat oleh hyperparameter importance (Gambar 5) yang menempatkan *learning rate* sebagai faktor dominan dan slice plot (Gambar 6) yang menunjukkan konsentrasi trial bernilai tinggi pada rentang *learning rate* moderat; pola tersebut sejalan dengan literatur bahwa skala pembaruan bobot menentukan kestabilan konvergensi dan kualitas generalisasi pada fine-tuning model pra-latih [2], [9]. Secara metodologis, penggunaan Optuna-TPE [7] juga dapat diposisikan lebih efisien daripada eksplorasi acak murni, yang dalam literatur optimasi hiperparameter dipaparkan sebagai baseline kuat namun tidak memanfaatkan informasi performa historis seperti pendekatan berbasis model [21].

3.2.3 Perbaikan pada Konsistensi BIO Dan Boundary Entitas (Analisis Per Label)

Analisis per label pada Tabel 4 menunjukkan bahwa peningkatan yang lebih menonjol terjadi pada label lanjutan I-* (misalnya I-LOC, I-ORG, I-PER) pada Optuna dibanding baseline, yang merefleksikan perbaikan konsistensi span entitas dalam skema BIO/IOB [1], [10], [24]. Hal ini penting karena label O mendominasi jumlah token (support tinggi), sehingga metrik agregat berpotensi "terangkat" oleh keberhasilan prediksi non-entitas; oleh karena itu, evaluasi perlu dibaca hingga tingkat label/entitas sebagaimana disajikan pada Tabel 4 [1], [10]. Bukti ini semakin kuat ketika dikaitkan dengan confusion matrix: Gambar 7 (baseline) dan Gambar 8 (Optuna) menunjukkan bahwa kesalahan dominan terkonsentrasi di sekitar batas span (token entitas diprediksi sebagai O atau pergeseran label B↔I), bukan pertukaran tipe entitas secara masif. Dengan demikian, peningkatan pada label I-* pada Tabel 4 selaras dengan berkurangnya kesalahan boundary yang tampak pada struktur kesalahan di confusion matrix, sehingga optimasi hiperparameter dapat dinilai memperbaiki aspek substantif ekstraksi entitas, bukan sekadar meningkatkan metrik agregat [24]. Pada konteks data media sosial, variasi

Copyright © 2026 The **Bambang Siswanto**, Page 231
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9545
Hal 221-233
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

ejaan dan noise umumnya memperbesar risiko kesalahan boundary; oleh karena itu, perbaikan konsistensi span pada label I-* menjadi indikator relevan terhadap peningkatan kualitas ekstraksi pada dataset Twitter/X [17].

3.2.4 Stabilitas Pelatihan dan Reprodusibilitas (Multi-Seed)

Hasil multi-seed pada tabel 5 menunjukkan Optuna meningkatkan rerata metrik sekaligus memperlihatkan pola sebaran yang lebih terkendali dibanding baseline seperti dalam Gambar 9. Temuan ini penting karena literatur menegaskan bahwa fine-tuning model pra-latih dapat menghasilkan variansi kinerja akibat inisialisasi bobot, urutan data, dan dinamika optimisasi, sehingga pelaporan lintas seed diperlukan untuk mengurangi bias "seed tertentu" dan memperkuat reprodusibilitas [2], [9]. Secara statistik, perbedaan performa multi-seed yang diringkas pada Tabel 6 menunjukkan paired t-test signifikan, sedangkan Wilcoxon berada sedikit di atas ambang signifikansi 0,05; perbedaan ini wajar pada ukuran sampel kecil karena uji non-parametrik dapat memiliki *power* yang lebih rendah dan *p-value* yang diskret. Oleh karena itu, interpretasi stabilitas lebih tepat dilakukan dengan mempertimbangkan konsistensi arah peningkatan antar-seed, bukti visual penyempitan sebaran (Gambar 9), serta bukti parametrik kuat dari paired t-test secara bersamaan [2], [9].

3.2.5 Robustness dan Generalisasi Lintas Partisi (K-Fold Cross-Validation)

Hasil K-Fold Cross-Validation sebagaimana tabel 7 menunjukkan performa yang tinggi dan stabil pada lima pembagian data, dengan rentang F1 yang sempit antar-fold. Secara metodologis, K-Fold CV merupakan evaluasi yang lebih ketat dibanding satu train–test split karena mengurangi ketergantungan pada komposisi data uji tertentu, sehingga estimasi performa lebih robust [21]. Dalam konteks NER yang distribusi labelnya tidak seimbang—terutama dominasi label O yang dapat memengaruhi metrik agregat, konsistensi antar-fold pada Tabel 7 memperkuat klaim bahwa peningkatan kinerja Optuna bersifat generalizable, bukan artefak dari pemilihan split tunggal [1], [10]. Temuan ini juga relevan untuk data Twitter/X yang heterogen, karena variasi gaya bahasa dan distribusi entitas dapat berubah antar-partisi; stabilitas antar-fold menjadi indikator bahwa konfigurasi Optuna lebih tahan terhadap variasi tersebut [17].

3.2.6 Implikasi Dan Batasan

Secara keseluruhan, rangkaian hasil dan pembahasan menunjukkan bahwa optimasi hiperparameter berbasis Optuna-TPE meningkatkan kinerja IndoBERT secara agregat, memperbaiki kualitas ekstraksi pada level boundary/span BIO, serta meningkatkan stabilitas dan robustness lintas partisi [2], [7], [9], [21]. Namun, interpretasi temuan tetap perlu mempertimbangkan keterbatasan yang melekat pada domain dataset (Twitter/X), distribusi label, serta skema anotasi, sehingga generalisasi ke domain lain sebaiknya divalidasi melalui evaluasi pada korpus berbeda atau perbandingan dengan model adaptasi domain media sosial pada penelitian lanjutan [10], [17].

4. KESIMPULAN

Penelitian ini mengevaluasi *fine-tuning* IndoBERT untuk Named Entity Recognition (NER) berbahasa Indonesia pada media sosial Twitter/X serta menguji dampak optimasi hiperparameter berbasis Optuna. Hasil eksperimen menunjukkan bahwa Optuna meningkatkan kinerja model dibandingkan baseline pada metrik evaluasi utama, sehingga menegaskan pentingnya penalaan hiperparameter yang sistematis untuk meningkatkan kemampuan generalisasi pada data media sosial yang heterogen. Selain meningkatkan kinerja, Optuna juga memperbaiki kualitas prediksi pada level entitas, khususnya terkait ketepatan batas dan konsistensi span pada skema BIO. Temuan ini mengindikasikan bahwa optimasi hiperparameter tidak hanya menaikkan skor, tetapi juga meningkatkan ketelitian ekstraksi entitas pada teks pendek. Dari sisi stabilitas, evaluasi multi-seed dan K-Fold *Cross-Validation* menunjukkan bahwa Optuna lebih stabil dan robust terhadap variasi inisialisasi maupun pembagian data. Kontribusi penelitian ini adalah menyediakan bukti empiris bahwa optimasi hiperparameter berbasis Optuna dapat meningkatkan kinerja sekaligus memperkuat stabilitas *fine-tuning* IndoBERT untuk NER Bahasa Indonesia pada domain Twitter/X. Ke depan, penelitian dapat diperluas melalui evaluasi lintas domain dan eksplorasi adaptasi domain untuk meningkatkan ketahanan model terhadap variasi bahasa yang lebih luas.

REFERENCES

[1] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: 10.1109/TKDE.2020.2981314.

[2] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. A. Smith, "Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping," *arXiv preprint arXiv:2002.06305*, 2020.

[3] M. B. Shishehgarkhaneh, R. C. Moehler, Y. Fang, A. A. Hijazi, and H. Aboutorab, "Transformer-Based Named Entity Recognition in Construction Supply Chain Risk Management in Australia," *IEEE Access*, vol. 12, no. March, pp. 41829–41851, 2024, doi: 10.1109/ACCESS.2024.3377232.

[4] S. O. Khairunnisa, "Dataset Enhancement and Multilingual Transfer for Named Entity Recognition in the Indonesian Language Dataset Enhancement and Multilingual Transfer for Named," vol. 22, no. 6, 2026, doi: 10.1145/3592854.

[5] F. Koto and T. Baldwin, "IndoLEM and IndoBERT : A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," pp. 757–770, 2020.

[6] T. Aprilah, D. R. Ignatius, M. Setiadi, and W. Herowati, "Enhancing Aspect-Based Sentiment Analysis via Hugging Face Fine-Tuned IndoBERT," vol. 9, no. 6, pp. 3821–3830, 2025.

Copyright © 2026 The **Bambang Siswanto**, Page 232
This Journal is licensed under a Creative Commons Attribution 4.0 International License

[7] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 2623–2631, 2019, doi: 10.1145/3292500.3330701.
[8] M. Khaldi, Z. Alilat, H. Bendoubba, and N. Mahammed, "Hyperparameter Optimization for Malicious URL Detection : Leveraging Optuna and Random Search in Machine Learning and Deep Learning Models 1 Introduction Related research," vol. 49, pp. 57–68, 2025.
[9] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the Stability of Fine-Tuning BERT: Misconceptions, Explanations, and Strong Baselines," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.
[10] B. Jehangir, S. Radhakrishnan, and R. Agarwal, "A survey on Named Entity Recognition — datasets, tools, and methodologies," *Nat. Lang. Process. J.*, vol. 3, p. 100017, 2023, doi: https://doi.org/10.1016/j.nlp.2023.100017.
[11] "Bio Tagging Ner Bahasa Indonesia (from twitter/X)." Accessed: Dec. 01, 2025. [Online]. Available: https://www.kaggle.com/datasets/gibranfktian/bio-tagging-ner-bahasa-indonesia-crawl-twitterx/data
[12] H. Liu, G. Jun, and Y. Zheng, "Chinese named entity recognition model based on BERT," vol. 06021, pp. 1–8, 2021.
[13] Y. Ling, Z. Ma, X. Dong, and X. Weng, "A deep learning approach for robust traffic accident information extraction from online chinese news," no. February, pp. 1847–1862, 2024, doi: 10.1049/itr2.12493.
[14] E. T. Luthfi, Z. Izzah, M. Yusoh, and B. M. Aboobaider, "BERT based Named Entity Recognition for Automated Hadith Narrator Identification," vol. 13, no. 1, pp. 604–611, 2022.
[15] B. S. Jati and M. N. Rizal, "Multilingual Named Entity Recognition Model for Indonesian Health Insurance Question Answering System," pp. 180–184, doi: 10.1109/ICOIACT50329.2020.9332027.
[16] Z. Liu, F. Jiang, Y. Hu, C. Shi, P. Fung, and A. Group, "NER-BERT: A Pre-trained Model for Low-Resource Entity Tagging," 2020.
[17] W. Liu and X. Cui, "applied sciences Improving Named Entity Recognition for Social Media with Data Augmentation," 2023.
[18] K. Kamdan, M. P. Anugrah, M. J. Almutaali, R. Ramdani, and I. L. Kharisma, "Performance Analysis of IndoBERT for Detection of Online Gambling Promotion in YouTube Comments †," 2025.
[19] B. Wilie *et al.*, "IndoNLU : Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," pp. 843–857, 2020.
[20] 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.
[21] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.
[22] T. Wolf *et al*., "Transformers: State-of-the-Art Natural Language Processing," in *Proc. EMNLP (System Demonstrations)*, 2020.
[23] S. Latisha, S. Favian, and D. Suhartono, "Criminal Court Judgment Prediction System Built on Modified BERT Models," vol. 15, no. 2, 2024, doi: 10.12720/jait.15.2.288-298.
[24] V. Yadav and S. Bethard, "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models," pp. 2145–2158, 2018.

This Journal is licensed under a Creative Commons Attribution 4.0 International License