E-ISSN 2963-9174

Indonesian Journal of Computer Science Research

Journal page: https://subset.id/index.php/IJCSR

# Implementasi Model IndoBERT pada *Named Entity Recognition* untuk Identifikasi Entitas dalam Cerita Rakyat Hikayat Wayang Arjuna & Purusara

# *Implementation of the IndoBERT Model on Named Entity Recognition for Entity Identification in the Folklore of the Tale of Wayang Arjuna & Purusara*

**Danda Bagaskoro[1,a)], Erna Daniati [2,b)], Rina Firliana [3,c)]**

[1, 2, 3)] Program Studi Sistem Informasi , Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri Jl. KH. Achmad Dahlan No.76, Kediri 64112, Indonesia

[a)] *Corresponding author:* Erna Daniati. Email: *ernadaniati@unpkediri.ac.id*
Received: 2 Juli 2026; Accepted: 29 Juli 2026

K E Y W O R DS

**Kata kunci:**
*Named Entity Recognition*
*IndoBERT*
*Cerita Rakyat*
*Teks*

***Keywords:***
*Named Entity Recognition*
*IndoBERT*
*Folklore*
*Text*

A B S T R A K

**Abstrak.** Perkembangan Natural Language Processing (NLP) mendorong pemanfaatan teknologi Named Entity Recognition (NER) untuk mengekstraksi informasi penting dari data teks secara otomatis. Penelitian ini bertujuan menerapkan model IndoBERT untuk mengenali entitas bernama pada teks sastra klasik Indonesia, yaitu Hikayat Wayang Arjuna & Purusara. Penelitian ini difokuskan pada identifikasi tiga kategori entitas, yaitu *Person* (PER), *Location* (LOC), dan *Organization* (ORG). Dataset disusun melalui proses anotasi menggunakan skema BIO *(Beginning, Inside, Outside)*, kemudian diproses melalui tahapan preprocessing, tokenisasi menggunakan IndoBERT Tokenizer, dan fine-tuning model IndoBERT. Evaluasi dilakukan menggunakan metrik precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa model IndoBERT memperoleh nilai precision sebesar 56,30%, recall sebesar 60,55%, dan F1-score sebesar 58,35%. Berdasarkan evaluasi setiap kategori entitas, kategori Location (LOC) memperoleh performa terbaik dengan nilai F1-score sebesar 76,07%, diikuti oleh Person (PER) sebesar 64,67%, sedangkan Organization (ORG) memperoleh nilai 44,96%. Hasil tersebut menunjukkan bahwa IndoBERT cukup efektif dalam mengenali entitas pada teks sastra klasik berbahasa Indonesia serta berpotensi mendukung proses ekstraksi informasi, digitalisasi, dan pelestarian cerita rakyat berbasis teknologi NLP.

***Abstract.*** *The development of Natural Language Processing (NLP) has encouraged the use of Named Entity Recognition (NER) technology to extract important information from text data automatically. This research aims to apply the IndoBERT model to recognize named entities in classic Indonesian literary texts, namely Hikayat Wayang Arjuna & Purusara. The research focused on the identification of three categories of entities, namely Person (PER), Location (LOC), and Organization (ORG). The dataset is compiled through an annotation process using the BIO (Beginning, Inside, Outside) scheme, then processed through the preprocessing stage, tokenization using the IndoBERT Tokenizer, and fine-tuning the IndoBERT model. Evaluation was conducted using precision, recall, and F1-score metrics. The results showed that the IndoBERT model obtained a precision score of 56.30%, recall of 60.55%, and an F1-score of 58.35%. Based on the evaluation of each entity category, the Location (LOC) category*

*obtained the best performance with an F1-score of 76.07%, followed by Person (PER) of 64.67%, while Organization (ORG) obtained a score of 44.96%. These results show that IndoBERT is quite effective in recognizing entities in classic literary texts in Indonesian and has the potential to support the process of information extraction, digitization, and preservation of folklore based on NLP technology.*

## 1. Pendahuluan

Kemajuan kecerdasan buatan telah mendorong perkembangan pesat dalam bidang Natural Language Processing (NLP), khususnya pada tugas Named Entity Recognition (NER) [1]. NER merupakan teknik untuk mengidentifikasi dan mengklasifikasikan entitas penting dalam teks seperti nama orang, lokasi, dan organisasi. Dalam konteks bahasa Indonesia, penerapan NER pada teks sastra klasik masih menghadapi tantangan besar akibat penggunaan bahasa arkais, istilah budaya yang kompleks, serta struktur kalimat yang berbeda dari bahasa modern. Salah satu teks yang memiliki karakteristik tersebut adalah Hikayat Wayang Arjuna & Purusara. Model berbasis transformer seperti IndoBERT memiliki kemampuan memahami konteks bahasa secara bidirectional sehingga berpotensi menghasilkan performa yang baik dalam tugas [2] Namun, performa model ini pada teks sastra klasik Indonesia belum banyak diteliti.Penelitian ini bertujuan menerapkan IndoBERT untuk mengenali entitas bernama pada teks hikayat serta mengevaluasi kemampuannya dalam memahami bahasa klasik dan istilah budaya pewayangan.

## 2. Tinjauan Pustaka

Natural Language Processing (NLP) merupakan cabang ilmu kecerdasan buatan yang berfokus pada pengolahan bahasa alami agar dapat dipahami oleh komputer. Salah satu tugas penting dalam NLP adalah *Named Entity Recognition (NER)*, yaitu proses mengidentifikasi dan mengklasifikasikan entitas bernama dalam suatu teks ke dalam kategori tertentu, seperti *Person (PER)***,** *Location (LOC)***,** dan *Organization (ORG)* atau kategori lain sesuai kebutuhan penelitian. Teknologi NER banyak dimanfaatkan dalam ekstraksi informasi, sistem temu kembali informasi, analisis dokumen, serta berbagai aplikasi pengolahan teks otomatis karena mampu mengubah informasi tidak terstruktur menjadi data yang lebih mudah dianalisis [3]. Perkembangan model berbasis *Transformer* telah meningkatkan performa pada berbagai tugas NLP, termasuk NER. Salah satu model yang banyak digunakan pada penelitian berbahasa Indonesia adalah *IndoBERT*, yaitu model bahasa yang telah melalui proses *pre-training* menggunakan korpus bahasa Indonesia dalam jumlah besar. Model ini mampu menghasilkan representasi kontekstual yang lebih baik dibandingkan dengan metode konvensional, sehingga banyak diterapkan pada klasifikasi teks, analisis sentimen, dan pengenalan entitas bernama. Dalam penelitian ini, IndoBERT dimanfaatkan sebagai model dasar (*pre-trained model*) yang kemudian disesuaikan untuk tugas klasifikasi token pada proses Named Entity Recognition [4]. Objek penelitian berupa cerita rakyat Hikayat Wayang Arjuna & Purusara dipilih karena memiliki karakteristik bahasa yang berbeda dengan teks modern. Cerita rakyat umumnya menggunakan kosakata arkais, istilah budaya, variasi ejaan, serta struktur kalimat yang lebih kompleks, sehingga menjadi tantangan dalam proses pengenalan entitas secara otomatis. Oleh karena itu, diperlukan model yang mampu memahami konteks bahasa secara mendalam agar proses identifikasi entitas dapat dilakukan dengan lebih akurat[5].

Agar model dapat beradaptasi dengan karakteristik data penelitian, dilakukan proses *fine-tuning* terhadap IndoBERT menggunakan dataset Named Entity Recognition berbahasa Indonesia yang telah diberi label menggunakan skema *BIO (Beginning, Inside, Outside)***.** Melalui proses tersebut, parameter model diperbarui sehingga mampu mengenali pola entitas sesuai dengan domain penelitian. Evaluasi performa model dilakukan menggunakan metrik *precision***,** *recall*, dan *F1-score*, yang merupakan metrik standar dalam pengukuran kinerja model NER [6]. Beberapa penelitian terdahulu menunjukkan bahwa model berbasis Transformer memberikan performa yang lebih baik dibandingkan dengan metode pembelajaran mesin konvensional pada berbagai tugas Named Entity Recognition. Penelitian mengenai IndoBERT menunjukkan bahwa model tersebut mampu meningkatkan kinerja pada berbagai tugas NLP berbahasa Indonesia, sedangkan penelitian lain membuktikan efektivitas model berbasis BERT pada pengenalan entitas di berbagai domain. Berdasarkan hasil penelitian sebelumnya, IndoBERT dipilih dalam penelitian ini karena memiliki kemampuan memahami konteks bahasa Indonesia secara lebih baik, sehingga diharapkan mampu mengenali entitas pada teks sastra klasik, khususnya *Hikayat Wayang Arjuna & Purusara.*

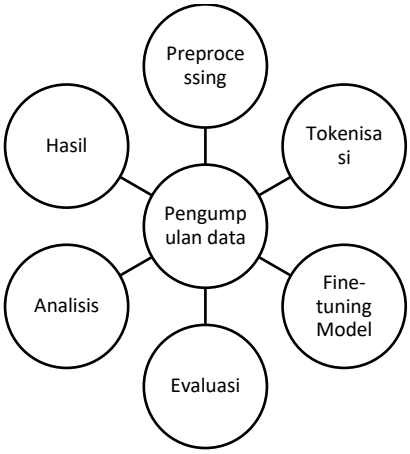
## 3. Metode Penelitian

Model IndoBERT yang digunakan merupakan model indobenchmark/indobert-base-p1 yang telah melalui proses pre-training menggunakan korpus bahasa Indonesia. Arsitektur model terdiri atas beberapa lapisan Transformer Encoder yang memanfaatkan mekanisme *multi-head self-attention* sehingga mampu menghasilkan representasi kontekstual setiap token. Pada penelitian ini, lapisan keluaran (*classification layer*) disesuaikan dengan jumlah label BIO yang digunakan sehingga setiap token dapat diklasifikasikan ke dalam kategori B-PER,

I-PER, B-LOC, I-LOC, B-ORG, I-ORG, maupun O. Selanjutnya, dilakukan fine-tuning model menggunakan dataset penelitian sehingga parameter model mampu beradaptasi terhadap karakteristik bahasa pada cerita rakyat Hikayat Wayang Arjuna & Purusara.

### *3.1. Desain Penelitian*

Adapun alur eksperimen pada penelitian disajikan pada Gambar 1.




**Gambar 1.** Alur penelitian

Berdasarkan Gambar 1**,** penelitian ini menggunakan metode eksperimen dengan pendekatan kuantitatif untuk mengevaluasi kemampuan model IndoBERT dalam melakukan tugas Named Entity Recognition (NER) pada teks cerita rakyat *Hikayat Wayang Arjuna & Purusara*. Tahapan penelitian dimulai dari pengumpulan dataset, persiapan data, pelabelan entitas, tokenisasi, proses *fine-tuning*, evaluasi model, hingga pengujian menggunakan kalimat-kalimat yang diambil dari cerita rakyat.

### *3.2. Dataset Penelitian*

Dataset diperoleh dari hasil anotasi teks Hikayat Wayang Arjuna & Purusara menggunakan skema BIO (Beginning, Inside, Outside). Seluruh data dipersiapkan dalam format CoNLL sehingga dapat digunakan pada proses token classification menggunakan model IndoBERT. Dataset kemudian dibagi menjadi data pelatihan sebanyak 1.464 data, data validasi sebanyak 367 data, dan data pengujian sebanyak 509 data untuk memastikan proses pelatihan dan evaluasi dilakukan secara objektif. [7].

### *3.3. Pelabelan Entitas*

Penelitian ini menggunakan skema pelabelan BIO (Beginning, Inside, Outside), yang merupakan salah satu metode standar dalam tugas *Named Entity Recognition*. Pada skema ini, setiap token diberikan label sesuai dengan kategori entitas yang dimilikinya, yaitu B untuk token pertama suatu entitas, I untuk token lanjutan dalam entitas yang sama, dan O untuk token yang tidak termasuk ke dalam entitas mana pun. Label yang digunakan meliputi *B-PER* dan *I-PER* untuk entitas *Person*, *B-LOC* dan *I-LOC* untuk entitas *Location*, serta *B-ORG* dan *I-ORG* untuk entitas *Organization*. Contoh penerapan skema BIO ditunjukkan pada Tabel 3, di mana token *Rajuna* diberi label *B-PER*, sedangkan *Ngamarta* diberi label *B-LOC*. Penggunaan skema BIO membantu model mengenali batas awal maupun batas lanjutan suatu entitas sehingga meningkatkan ketepatan proses klasifikasi.

### *3.4. Implementasi Model IndoBERT*

Model yang digunakan pada penelitian ini adalah indobenchmark/indobert-base-p1 yang tersedia melalui pustaka Hugging Face Transformers. Model tersebut merupakan model bahasa berbasis arsitektur *Bidirectional Encoder Representations from Transformers (BERT)* yang telah melalui proses *pre-training* menggunakan korpus bahasa Indonesia dalam jumlah besar. Dengan proses tersebut, IndoBERT mampu memahami hubungan antarkata berdasarkan konteks kalimat secara dua arah (*bidirectional*), sehingga memiliki kemampuan yang baik dalam berbagai tugas *Natural Language Processing* (NLP), termasuk *Named Entity Recognition (NER).* Pada penelitian ini, model IndoBERT diadaptasi untuk melakukan token *classification*, yaitu memberikan label pada setiap token berdasarkan kategori entitas menggunakan skema BIO (Beginning, Inside, Outside). Selanjutnya, model menjalani proses *fine-tuning* menggunakan dataset yang telah dianotasi sehingga parameter model dapat menyesuaikan dengan karakteristik bahasa pada teks *Hikayat Wayang Arjuna & Purusara*. Implementasi dilakukan menggunakan bahasa pemrograman Python dengan framework PyTorch serta pustaka Hugging Face Transformers. Penambahan lapisan klasifikasi (*classification layer*) pada keluaran encoder memungkinkan model memprediksi label setiap token sesuai kategori entitas Person (PER), Location (LOC), dan Organization (ORG). Adapun representasi kontekstual setiap token pada *IndoBERT* disajikan pada formula (1).

$$H = Tranformer\ (\mathrm{X} + \mathrm{P} + \mathrm{S}) \quad (1)$$

Di mana *H* = representasi kontekstual hasil keluaran encoder Transformer, *X* = embedding token masukan (token embedding), P = positional embedding yang menunjukkan posisi token dalam kalimat, S = segment embedding yang membedakan pasangan kalimat, *Transformer = encoder* BERT yang memanfaatkan mekanisme self-attention untuk memperoleh representasi kontekstual setiap token. Melalui mekanisme tersebut, setiap token dapat memperhatikan hubungan dengan seluruh token lainnya sehingga model mampu memahami konteks kalimat secara menyeluruh. Pendekatan ini menjadikan IndoBERT efektif dalam mengenali entitas pada teks berbahasa Indonesia, termasuk teks sastra klasik yang memiliki karakteristik bahasa berbeda dengan bahasa modern.

### *3.5. Tokenisasi dan Fine-Tuning*

Sebelum proses pelatihan dilakukan, seluruh kalimat diproses menggunakan *IndoBERT Tokenizer* untuk mengubah teks menjadi token yang dapat dipahami oleh model. Karena satu kata dapat dipecah menjadi beberapa *sub-token*, dilakukan proses *label alignment* agar setiap token tetap memiliki label yang sesuai dengan hasil anotasi. Setelah proses tokenisasi selesai, dilakukan *fine-tuning* terhadap model IndoBERT menggunakan data pelatihan. Proses pelatihan dilakukan dengan parameter berupa model *indobenchmark/indobert-base-p1***,** *epoch* sebanyak 40**,** *learning rate* sebesar $3 \times 10^{-5}$**,** *batch size* pelatihan 8**,** *batch size* validasi 32, *weight decay* 0,01, *warmup ratio* 0,2**,** serta *optimizer* AdamW**.** Pemilihan parameter tersebut bertujuan menghasilkan proses pelatihan yang stabil dan memperoleh performa model yang optimal pada data validasi.

### *3.6. Evaluasi Model*

Evaluasi model dilakukan untuk mengukur kemampuan IndoBERT dalam mengenali entitas bernama pada dataset pengujian. Pengukuran performa menggunakan tiga metrik utama, yaitu *Precision, Recall*, dan *F1-Score*, yang merupakan metrik standar pada tugas *Named Entity Recognition (NER)*. Ketiga metrik tersebut dihitung berdasarkan jumlah prediksi yang benar (*True Positive*), prediksi salah (*False Positive*), dan entitas yang tidak berhasil dikenali (*False Negative*). Evaluasi model pada penelitian ini menggunakan precision, recall dan F1-Score, yang disajikan pada (2), (3), dan (4).

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

Di mana TP (True Positive) = jumlah entitas yang berhasil diprediksi dengan benar, FP (False Positive) = jumlah entitas yang diprediksi sebagai entitas, tetapi sebenarnya bukan entitas. Precision menunjukkan tingkat ketepatan model dalam memberikan prediksi terhadap suatu entitas. Nilai presisi yang tinggi menunjukkan bahwa sebagian besar hasil prediksi model merupakan entitas yang benar.

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

Di mana TP (True Positive) = jumlah prediksi benar, FN (False Negative) = jumlah entitas sebenarnya tetapi gagal dikenali oleh model. Recall menggambarkan kemampuan model dalam menemukan seluruh entitas yang terdapat pada data pengujian. Semakin tinggi nilai recall, semakin sedikit entitas yang terlewat oleh model.

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

Di mana Precision = tingkat ketepatan prediksi model, Recall = tingkat keberhasilan model dalam menemukan seluruh entitas. F1-Score merupakan rata-rata harmonik antara precision dan recall sehingga mampu memberikan gambaran performa model secara lebih seimbang. Nilai F1-Score digunakan sebagai indikator utama dalam mengevaluasi keberhasilan model NER karena mempertimbangkan ketepatan dan kelengkapan hasil prediksi secara bersamaan.

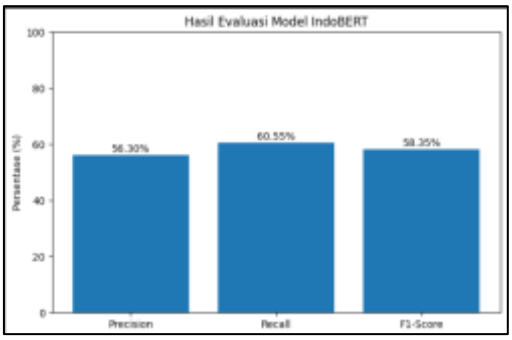
## 4. Pembahasan

Hasil penelitian menunjukkan bahwa model IndoBERT mampu melakukan tugas Named Entity Recognition pada teks Hikayat Wayang Arjuna & Purusara dengan nilai precision sebesar 56,30%, recall sebesar 60,55%, dan F1-score sebesar 58,35%. Nilai tersebut menunjukkan bahwa model telah mampu mengenali sebagian besar entitas yang terdapat pada teks sastra klasik berbahasa Indonesia, meskipun performanya masih dapat ditingkatkan. Dengan demikian, tujuan penelitian untuk menerapkan model IndoBERT dalam mengidentifikasi entitas bernama pada cerita rakyat berhasil dicapai karena model mampu melakukan klasifikasi terhadap entitas yang terdapat pada dataset penelitian.

Berdasarkan hasil evaluasi setiap kategori entitas, Location (LOC) memperoleh nilai F1-score tertinggi sebesar 76,07%, diikuti oleh Person (PER) sebesar 64,67%, sedangkan Organization (ORG) memperoleh nilai terendah sebesar 44,96%. Performa yang lebih tinggi pada kategori lokasi menunjukkan bahwa nama tempat dalam teks Hikayat Wayang Arjuna & Purusara memiliki pola penulisan yang relatif konsisten sehingga lebih mudah

dipelajari oleh model. Sebaliknya, kategori organisasi memiliki performa yang lebih rendah karena jumlah data latih pada kategori tersebut lebih sedikit serta memiliki karakteristik penamaan yang lebih bervariasi sehingga meningkatkan kemungkinan terjadinya kesalahan klasifikasi.

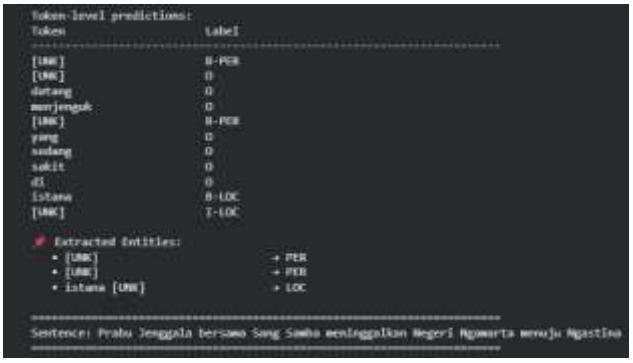
Hasil pelatihan juga menunjukkan bahwa nilai F1-score mengalami peningkatan selama proses fine-tuning, sedangkan training loss terus mengalami penurunan. Akan tetapi, peningkatan validation loss pada epoch-epoch akhir mengindikasikan adanya kecenderungan overfitting ringan. Kondisi tersebut menunjukkan bahwa model mulai menyesuaikan diri secara berlebihan terhadap data pelatihan sehingga kemampuan generalisasi terhadap data baru menjadi berkurang. Meskipun demikian, performa yang diperoleh masih menunjukkan bahwa IndoBERT mampu memanfaatkan representasi kontekstual bahasa Indonesia untuk mengenali entitas pada teks sastra klasik.




**Gambar 2.** Grafik Hasil Pelatihan Model IndoBERT

Grafik hasil pelatihan menunjukkan bahwa nilai training loss mengalami penurunan secara bertahap selama proses fine-tuning, yang mengindikasikan bahwa model semakin mampu mempelajari pola data pelatihan. Sementara itu, nilai validation loss cenderung stabil pada sebagian besar epoch, namun mengalami sedikit peningkatan pada epoch-epoch akhir. Kondisi tersebut menunjukkan adanya kecenderungan overfitting ringan, yaitu model mulai menyesuaikan diri secara berlebihan terhadap data pelatihan sehingga kemampuan generalisasi terhadap data baru sedikit menurun. Meskipun demikian, model masih mampu menghasilkan performa yang cukup baik dengan nilai Precision sebesar 56,30%, Recall 60,55%, dan F1-Score 58,35%, sehingga masih layak digunakan untuk proses identifikasi entitas pada teks Hikayat Wayang Arjuna & Purusara. [8].

Hasil penelitian ini menunjukkan bahwa model IndoBERT mampu melakukan tugas Named Entity Recognition pada teks sastra klasik berbahasa Indonesia dengan performa yang cukup baik. Temuan tersebut sejalan dengan penelitian [4] yang menyatakan bahwa model berbasis pre-trained language model mampu menghasilkan representasi kontekstual yang lebih baik dibandingkan pendekatan word embedding konvensional pada tugas NER bahasa Indonesia. Selain itu, penelitian [5] juga menunjukkan bahwa model berbasis deep learning memiliki kemampuan yang baik dalam mengenali entitas pada teks berbahasa Indonesia, meskipun performanya dipengaruhi oleh karakteristik dataset dan kualitas proses anotasi. Pada penelitian ini, tantangan utama berasal dari penggunaan bahasa sastra klasik yang mengandung kosakata arkais, istilah budaya, dan variasi penulisan yang lebih kompleks dibandingkan dengan teks modern [9].. Kondisi tersebut menyebabkan performa model pada kategori entitas tertentu, terutama Organization (ORG), masih lebih rendah dibandingkan kategori lainnya. Meskipun demikian, hasil penelitian ini memperlihatkan bahwa IndoBERT tetap memiliki kemampuan yang baik dalam memahami konteks bahasa Indonesia, sehingga berpotensi diterapkan sebagai pendekatan untuk ekstraksi informasi pada naskah budaya dan cerita rakyat.




**Gambar 3**. Hasil Pengujian Model IndoBERT pada Cerita Hikayat Wayang

Gambar 3 menunjukkan hasil pengujian model IndoBERT terhadap salah satu kalimat pada teks Hikayat Wayang Arjuna & Purusara. Model berhasil memberikan label entitas pada setiap token menggunakan skema BIO (Beginning, Inside, Outside). Token yang termasuk nama tokoh dikenali sebagai Person (PER), nama tempat dikenali sebagai Location (LOC), sedangkan nama organisasi atau kelompok dikenali sebagai Organization (ORG). Hasil ini menunjukkan bahwa model mampu melakukan identifikasi entitas secara otomatis meskipun teks yang digunakan merupakan teks sastra klasik yang memiliki karakteristik bahasa yang berbeda dengan bahasa Indonesia modern. Berdasarkan hasil pengujian tersebut, sebagian besar entitas berhasil dikenali sesuai dengan label sebenarnya. Model menunjukkan performa yang baik pada entitas Location (LOC) karena pola penulisan nama lokasi pada teks relatif konsisten. Sebaliknya, beberapa kesalahan prediksi masih ditemukan pada kategori Organization (ORG) akibat jumlah data pelatihan yang lebih sedikit serta variasi penulisan nama organisasi yang cukup beragam. Walaupun demikian, hasil eksperimen membuktikan bahwa model IndoBERT mampu menghasilkan prediksi yang cukup akurat dalam mengenali entitas bernama pada cerita rakyat Indonesia.

## 5. Kesimpulan

Penelitian ini berhasil mengimplementasikan model IndoBERT untuk melakukan tugas Named Entity Recognition (NER) pada teks cerita rakyat Hikayat Wayang Arjuna & Purusara. Seluruh tahapan penelitian telah dilaksanakan mulai dari penyusunan dataset berbahasa Indonesia yang diberi anotasi menggunakan skema BIO (Beginning, Inside, Outside), proses tokenisasi menggunakan IndoBERT Tokenizer, fine-tuning model, hingga evaluasi menggunakan metrik precision, recall, dan F1-score. Berdasarkan hasil evaluasi, model memperoleh nilai precision sebesar 56,30%, recall sebesar 60,55%, dan F1-score sebesar 58,35%, yang menunjukkan bahwa IndoBERT mampu mengenali entitas bernama pada teks sastra klasik berbahasa Indonesia dengan performa yang cukup baik. Hasil tersebut menunjukkan bahwa tujuan penelitian untuk menerapkan model IndoBERT dalam mengidentifikasi entitas pada cerita rakyat berhasil dicapai. Temuan penelitian menunjukkan bahwa performa model berbeda pada setiap kategori entitas. Kategori Location (LOC) memperoleh hasil terbaik dengan nilai F1-score sebesar 76,07%, diikuti oleh kategori Person (PER) sebesar 64,67%, sedangkan kategori Organization (ORG) memperoleh nilai F1-score sebesar 44,96%. Perbedaan performa tersebut mengindikasikan bahwa karakteristik data dan konsistensi pola penulisan entitas berpengaruh terhadap kemampuan model dalam melakukan klasifikasi. Nama lokasi pada teks memiliki pola yang lebih konsisten sehingga lebih mudah dikenali oleh model, sedangkan entitas organisasi memiliki variasi penulisan yang lebih tinggi sehingga tingkat kesalahan prediksi menjadi lebih besar. Selain itu, hasil pelatihan memperlihatkan adanya kecenderungan overfitting ringan pada epoch akhir, yang ditandai dengan meningkatnya validation loss meskipun training loss terus menurun. Kondisi tersebut menunjukkan bahwa optimasi parameter dan peningkatan kualitas data masih diperlukan agar kemampuan generalisasi model dapat ditingkatkan Secara keseluruhan, penelitian ini membuktikan bahwa IndoBERT memiliki potensi untuk diterapkan pada proses ekstraksi informasi otomatis dari teks sastra klasik Indonesia. Pemanfaatan model ini dapat mendukung digitalisasi dokumen budaya dengan membantu proses identifikasi tokoh, lokasi, dan entitas penting secara lebih cepat dibandingkan dengan analisis manual. Temuan ini juga memberikan kontribusi terhadap pengembangan penelitian Natural Language Processing (NLP) berbahasa Indonesia, khususnya pada penerapan Named Entity Recognition untuk naskah budaya dan cerita rakyat[10]. Penelitian selanjutnya mengembangkan model lain untuk mendeteksi teks.

## Daftar Pustaka


[1] S. Ferdiana, P. Wibowo, A. Famasya, and S. Basuki, "Named Entity Recognition in Medical Domain : A Systematic Literature Review," vol. 8, no. December, 2024.

[2] E. Daniati, A. Prasetya, W. Sakti, G. Irianto, and L. Hernandez, "Few-Shot-BERT-RNN Narrative Structure Analysis for Andersen ' s Stories," vol. 9, no. July, pp. 1773–1782, 2025.

[3] L. Zhao and L. Li, "A BERT based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts".

[4] M. Susanty, "Perbandingan Pre-trained Word Embedding dan Embedding Layer untuk Named-Entity Recognition Bahasa Indonesia," vol. 14, no. 2, pp. 247–257, 2021.

[5] A. H. Pradhana, E. Daniati, and M. Najibulloh, "Penerapan Bi-LSTM Untuk Named Entity Recognition Pada Teks Bahasa Indonesia," vol. 4, no. 2, pp. 96–106, 2025.

[6] M. Faruqziddan, E. Daniati, and M. N. Muzaki, "Pendekatan BERT Dalam Analisis Sentimen Terhadap Kominfo Di Media Sosial X," vol. 4, no. 2, pp. 107–116, 2025.

[7] X. Chen et al., "Good Visual Guidance Makes A Better Extractor : Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction," pp. 1607–1618, 2022.

[8] S. Feuerriegel et al., "Using natural language processing to analyse text data in behavioural science", doi: 10.1038/s44159-024-00392-z.

[9] E. Daniati et al., "Perbandingan Kinerja Algoritma SVM , LSTM , dan Fine- tuned IndoBERT dalam Analisis Sentimen Opini Masyarakat Indonesia terhadap Mobil Listrik," vol. 5, no. 1, pp. 1–13, 2026.

[10] L. Cui, Y. Wu, J. Liu, S. Yang, and Y. Zhang, "Template-Based Named Entity Recognition Using BART," pp. 1835–1845, 2021.

[11] W. Wang, S. Tian, H. Wang, and T. Lookman, "Applications of natural language processing and large language models in materials discovery," npj Comput. Mater., pp. 1–15, 2025, doi: 10.1038/s41524-025-01554-0.