

*NAMED ENTITY RECOGNITION (NER) PADA TEKS BERBAHASA  
INDONESIA DENGAN FINE-TUNING INDOBERT*

*Diajukan Untuk Menyusun Skripsi  
Di Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya*



Oleh:

Affandi Arrizal  
NIM: 09021182126027

**JURUSAN TEKNIK INFORMATIKA  
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA  
2024**

## LEMBAR PENGESAHAN SKRIPSI

*NAMED ENTITY RECOGNITION (NER) PADA TEKS BERBAHASA  
INDONESIA DENGAN FINE-TUNING INDOBERT*

Oleh:

Affandi Arrizal

NIM: 09021182126027

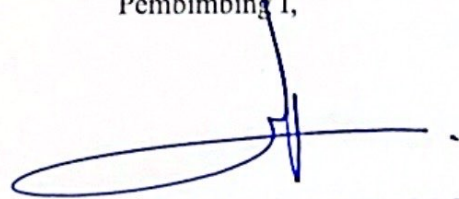
Indralaya, 23 Desember 2024

Mengetahui  
Ketua Jurusan Teknik Informatika,



Hadipurnawan Satria, Ph.D.  
NIP. 198004182020121001

Pembimbing I,

A handwritten signature in blue ink, consisting of a large loop and a vertical stroke, positioned above the name of the supervisor.

Dr. Abdiansah, S.Kom., M.Cs.  
NIP. 198410012009121005

## TANDA LULUS UJIAN KOMPREHENSIF

Pada hari senin tanggal 23 Desember 2024 telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Affandi Arrizal

NIM : 09021182126027

Judul : *Named Entity Recognition (NER)* pada Teks Berbahasa Indonesia dengan *Fine-Tuning IndoBERT*

Dan dinyatakan LULUS.

1. Ketua Penguji

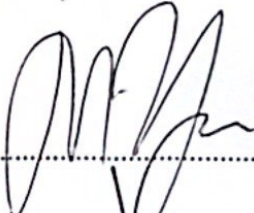
Desty Rodiah, S.Kom, M.T.  
NIP. 198912212020122011



.....

2. Penguji

Novi Yusliani, S.Kom, M.T.  
NIP. 198211082012122001



.....

3. Pembimbing

Dr. Abdiansah, S.Kom., M.Cs.  
NIP. 198410012009121005



.....

Mengetahui,  
Ketua Jurusan Teknik Informatika



Hadipurnayan Satria, S.Kom., M.Sc., Ph.D.  
NIP. 198004182020121001

## HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Affandi Arrizal

NIM : 09021182126027

Program Studi : Teknik Informatika

Judul Skripsi : *Named Entity Recognition (NER)* pada Teks Berbahasa Indonesia dengan *Fine-Tuning IndoBERT*

**Hasil Pengecekan *Software Turnitin*: 5%**

Menyertakan bahwa laporan tugas akhir saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan proyek ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan dari pihak manapun.



Indralaya, 30 Desember 2024



Affandi Arrizal

NIM. 09021182126027

## MOTTO DAN PERSEMBAHAN

Motto:

*“Tidak ada mimpi yang gagal, yang ada hanyalah mimpi yang tertunda. Jika sekiranya merasa gagal dalam mencapai mimpi, jangan khawatir. Mimpi-mimpi lain bisa diciptakan.”*

- Brando Franco Windah (Windah Basudara)

Kupersembahkan Karya Tulis ini kepada:

- Allah SWT
- Orang Tua
- Keluarga Besar
- Fakultas Ilmu Komputer
- Universitas Sriwijaya

## ***ABSTRACT***

*Textual document processing often requires the identification of important information such as names of people, places and institutions, but manual processes are time-consuming and inefficient. Named Entity Recognition (NER) is an automated solution for this task, but its application in Indonesian faces challenges such as dialectal variations and complex linguistic structures. This research applies a fine-tuning approach to the 'indobenchmark/indobert-base-p2' model with the NER-Grit dataset from a GitHub repository to tackle the challenges associated with the diverse linguistic structure of the Indonesian language. The model was tested using various parameter configurations to obtain optimal results. The evaluation showed that the best configuration with a learning rate of  $5e-6$ , batch size of 8, and 10 epochs resulted in an f1-score of 0.7420 (or 74,20%) and a loss value of 0.3673. This research contributes to the development of NER systems that are more adaptive to domain and dialect variations in Indonesian, thus supporting faster and more accurate information processing for various analytical needs and practical applications.*

***Keywords:*** *Named Entity Recognition, IndoBERT, Fine-tuning, Natural Language Processing*

## ABSTRAK

Pengolahan dokumen tekstual sering kali membutuhkan identifikasi informasi penting seperti nama orang, tempat, dan lembaga, namun proses manual memakan waktu dan kurang efisien. *Named Entity Recognition* (NER) menjadi solusi otomatis untuk tugas ini, tetapi penerapannya dalam bahasa Indonesia menghadapi tantangan seperti variasi dialek dan struktur linguistik yang kompleks. Penelitian ini menggunakan metodologi *fine-tuning* pada model 'indobenchmark/indobert-base-p2' dengan dataset NER-Grit dari repositori GitHub untuk menghadapi tantangan struktur linguistik bahasa Indonesia yang beragam. Model diuji menggunakan berbagai konfigurasi parameter untuk memperoleh hasil optimal. Evaluasi menunjukkan bahwa konfigurasi terbaik dengan *learning rate*  $5e-6$ , *batch size* 8, dan 10 *epoch* menghasilkan nilai *f1-score* 0,7420 (atau 74,20%) dan nilai *loss* 0,3673. Penelitian ini berkontribusi pada pengembangan sistem NER yang lebih adaptif terhadap variasi domain dan dialek dalam bahasa Indonesia, sehingga mendukung pengolahan informasi secara lebih cepat dan akurat untuk berbagai kebutuhan analitik dan aplikasi praktis.

**Kata Kunci:** *Named Entity Recognition*, IndoBERT, *Fine-tuning*, Pemrosesan Bahasa Alami

## KATA PENGANTAR

Puji dan syukur kepada Allah SWT yang telah memberikan nikmat sehat, iman dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Skripsi ini dibuat sebagai salah satu syarat untuk menyelesaikan Pendidikan program Strata-1 di Fakultas Ilmu Komputer Universitas Sriwijaya. Dalam proses pembuatan skripsi ini penulis menerima bimbingan, bantuan, semangat, maupun petunjuk dari banyak pihak. Oleh karena itu, penulis ingin menyampaikan terima kasih kepada:

1. Allah SWT atas rahmat dan nikmat-Nya sehingga, penulis dapat menyelesaikan skripsi ini dengan baik.
2. Kedua orang tua tercinta, Bapak Rubangi dan Ibu Muji Astuti, serta kakak tersayang Annisa Sabela Ningrum, yang senantiasa memberikan dukungan moril, materiil, dan doa dalam setiap langkah perjuangan penulis.
3. Bapak Hadipurnawan Satria, Ph.D. selaku ketua jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
4. Ibu Alvi Syahrini Utami, S.Si., M.Kom. selaku Dosen dan sekaligus pembimbing akademik yang telah memberikan banyak bantuan dan arahan kepada penulis selama perkuliahan.
5. Bapak Dr. Abdiansah, S.Kom., M.Cs. selaku Dosen Pembimbing skripsi yang telah memberi bimbingan, arahan serta semangat kepada penulis dalam menyelesaikan skripsi ini.

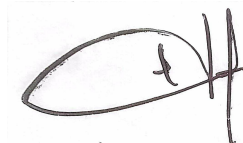


6. Seluruh dosen program studi serta admin Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
7. Seluruh staf Administrasi dan Pegawai Fakultas Ilmu Komputer yang telah membantu dalam urusan administrasi.
8. Rekan-rekan perjuangan penulis selama perkuliahan, Iman, Dzaky, Zidane, Efan, Wahyu, Anharul, Bima, Hafiz, Aditya, dan Yudha, yang selalu mendukung dan mengingatkan penulis dalam menyelesaikan skripsi ini.
9. Sahabat-sahabat penulis, “Sound0” dan “A-8 Supernova”, yang senantiasa memberikan dukungan dan keceriaan kepada penulis selama proses pengerjaan skripsi ini.
10. Pihak-pihak lain yang tidak dapat penulis sebutkan satu-persatu.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih banyak sekali kekurangan karena kurangnya pengalaman dan pengetahuan penulis. Oleh karena itu penulis mengharapkan saran dan kritik yang membangun guna kemajuan penelitian selanjutnya. Semoga tugas akhir ini dapat bermanfaat. Terima Kasih.

Indralaya, 23 Desember 2024

Penulis,

A handwritten signature in black ink, consisting of a large, stylized loop followed by several vertical strokes.

Affandi Arrizal

## DAFTAR ISI

<b>LEMBAR PENGESAHAN SKRIPSI</b> .....	i
<b>TANDA LULUS UJIAN KOMPREHENSIF</b> .....	ii
<b>HALAMAN PERNYATAAN</b> .....	iii
<b>MOTTO DAN PERSEMBAHAN</b> .....	iv
<b>ABSTRACT</b> .....	v
<b>ABSTRAK</b> .....	vi
<b>KATA PENGANTAR</b> .....	vii
<b>DAFTAR ISI</b> .....	ix
<b>DAFTAR GAMBAR</b> .....	xiii
<b>DAFTAR TABEL</b> .....	xv
<b>BAB I</b> .....	I-1
1.1    Pendahuluan .....	I-1
1.2    Latar Belakang Masalah .....	I-1
1.3    Rumusan Masalah .....	I-3
1.4    Tujuan Penelitian .....	I-4
1.5    Manfaat Penelitian .....	I-4
1.6    Batasan Masalah .....	I-5
1.7    Sistematika Penulisan .....	I-5
1.8    Kesimpulan .....	I-7
<b>BAB II</b> .....	II-1
2.1    Pendahuluan .....	II-1
2.2    Landasan Teori .....	II-1
2.2.1 <i>Natural Language Processing</i> .....	II-1
2.2.2 <i>Named Entity Recognition</i> .....	II-2
2.2.3    IndoBERT .....	II-3
2.2.4 <i>Fine-Tuning</i> .....	II-7

2.2.5	<i>WordPiece Tokenization</i> .....	II-9
2.2.6	<i>Confusion Matrix</i> .....	II-12
2.2.7	<i>Confidence Score</i> .....	II-16
2.2.8	<i>Rational Unified Process (RUP)</i> .....	II-17
2.3	Penelitian Terdahulu .....	II-19
2.4	Kesimpulan.....	II-22
<b>BAB III</b>	.....	<b>III-1</b>
3.1	Pendahuluan .....	III-1
3.2	Pengumpulan Data.....	III-1
3.2.1	Jenis dan Sumber Data .....	III-1
3.2.2	Metode Pengumpulan Data.....	III-2
3.3	Tahapan Penelitian.....	III-5
3.3.1	Mengumpulkan Data.....	III-6
3.3.2	Menentukan Kerangka Kerja Penelitian .....	III-6
3.3.3	Menentukan Kriteria Pengujian .....	III-16
3.3.4	Menentukan Format Data Pengujian.....	III-17
3.3.5	Menentukan Alat Bantu Pengujian.....	III-18
3.3.6	Melakukan Pengujian Penelitian.....	III-19
3.3.7	Melakukan Analisis dan Menarik Kesimpulan Penelitian .....	III-19
3.4	Metode Pengembangan Perangkat Lunak .....	III-20
3.5	Manajemen Proyek Penelitian.....	III-21
3.6	Kesimpulan.....	III-25
<b>BAB IV</b>	.....	<b>IV-1</b>
4.1	Pendahuluan .....	IV-1
4.2	Fase Insepsi .....	IV-1
4.2.1	Pemodelan Bisnis .....	IV-1
4.2.2	Kebutuhan Sistem .....	IV-2
4.2.3	Analisis dan Perancangan .....	IV-2

4.3	Fase Elaborasi.....	IV-7
4.3.1	Pemodelan Bisnis .....	IV-7
4.3.2	Kebutuhan Sistem .....	IV-9
4.3.3	Analisis dan Perancangan .....	IV-10
4.4	Fase Konstruksi .....	IV-14
4.4.1	Kebutuhan Sistem .....	IV-14
4.4.2	Implementasi.....	IV-15
4.5	Fase Transisi.....	IV-16
4.5.1	Pemodelan Bisnis.....	IV-16
4.5.2	Rencana Pengujian.....	IV-17
4.5.3	Implementasi.....	IV-17
4.6	Kesimpulan.....	IV-19
<b>BAB V</b>	.....	V-1
5.1	Pendahuluan .....	V-1
5.2	Hasil Penelitian.....	V-1
5.2.1	Konfigurasi Percobaan.....	V-1
5.2.2	Hasil Pengujian Konfigurasi I.....	V-3
5.2.3	Hasil Pengujian Konfigurasi II.....	V-5
5.2.4	Hasil Pengujian Konfigurasi III .....	V-7
5.2.5	Hasil Pengujian Konfigurasi IV .....	V-9
5.2.6	Hasil Pengujian Konfigurasi V .....	V-11
5.2.7	Hasil Pengujian Konfigurasi VI.....	V-13
5.2.8	Hasil Pengujian Konfigurasi VII.....	V-15
5.2.9	Hasil Pengujian Konfigurasi VIII .....	V-17
5.3	Analisis Hasil Penelitian.....	V-19
5.4	Kesimpulan.....	V-21
<b>BAB VI</b>	.....	VI-1
6.1	Pendahuluan .....	VI-1

6.2	Kesimpulan.....	VI-1
6.3	Saran.....	VI-2
<b>DAFTAR PUSTAKA.....</b>		<b>xiv</b>

## DAFTAR GAMBAR

<b>Gambar II - 1.</b> Arsitektur BERT .....	II-5
<b>Gambar II - 2.</b> Masked Language Model .....	II-6
<b>Gambar II - 3.</b> Next Sentence Prediction .....	II-7
<b>Gambar II - 4.</b> Dimensi RUP.....	II-18
<b>Gambar III - 1.</b> Tahapan Penelitian.....	III-5
<b>Gambar III - 2.</b> Alur Tahapan Penelitian.....	III-6
<b>Gambar III - 3.</b> Proses Tokenisasi Awal (WordPiece Tokenizer).....	III-8
<b>Gambar III - 4.</b> Penambahan Token Khusus .....	III-9
<b>Gambar III - 5.</b> Konversi Token menjadi ID Numerik .....	III-10
<b>Gambar III - 6.</b> Attention Mask.....	III-10
<b>Gambar III - 7.</b> Positional Embedding.....	III-11
<b>Gambar III - 8.</b> Penambahan Token Khusus & Pemecahan Kata .....	III-12
<b>Gambar III - 9.</b> Penyesuaian Label Pasca-Tokenisasi.....	III-13
<b>Gambar IV - 1.</b> Use Case Diagram.....	IV-3
<b>Gambar IV - 2.</b> Wireframe 1 .....	IV-8
<b>Gambar IV - 3.</b> Wireframe 2 .....	IV-9
<b>Gambar IV - 4.</b> Activity Diagram.....	IV-11
<b>Gambar IV - 5.</b> Sequence Diagram Memuat Data .....	IV-12
<b>Gambar IV - 6.</b> Sequence Diagram Menampilkan Hasil Prediksi.....	IV-13
<b>Gambar IV - 7.</b> Class Diagram .....	IV-14
<b>Gambar IV - 8.</b> Implementasi Antarmuka 1 .....	IV-15
<b>Gambar IV - 9.</b> Implementasi Antarmuka 2 .....	IV-16
<b>Gambar V - 1.</b> Grafik Pengujian Konfigurasi I .....	V-3
<b>Gambar V - 2.</b> Confusion Matrix Konfigurasi I .....	V-4
<b>Gambar V - 3.</b> Grafik Hasil Percobaan dan Validasi Konfigurasi II .....	V-5
<b>Gambar V - 4.</b> Confusion Matrix Konfigurasi II.....	V-6

<b>Gambar V - 5.</b> Grafik Pengujian Konfigurasi III.....	V-7
<b>Gambar V - 6.</b> Confusion Matrix Konfigurasi III.....	V-8
<b>Gambar V - 7.</b> Grafik Hasil Percobaan dan Validasi Konfigurasi IV .....	V-9
<b>Gambar V - 8.</b> Confusion Matrix Konfigurasi IV .....	V-10
<b>Gambar V - 9.</b> Grafik Hasil Percobaan dan Validasi Konfigurasi V .....	V-11
<b>Gambar V - 10.</b> Confusion Matrix Konfigurasi V.....	V-12
<b>Gambar V - 11.</b> Grafik Hasil Percobaan dan Validasi Konfigurasi VI .....	V-13
<b>Gambar V - 12.</b> Confusion Matrix Konfigurasi VI .....	V-14
<b>Gambar V - 13.</b> Grafik Hasil Percobaan dan Validasi Konfigurasi VII.....	V-15
<b>Gambar V - 14.</b> Confusion Matrix Konfigurasi VII .....	V-16
<b>Gambar V - 15.</b> Grafik Hasil Percobaan dan Validasi Konfigurasi VIII .....	V-17
<b>Gambar V - 16.</b> Confusion Matrix Konfigurasi VIII.....	V-18
<b>Gambar V - 13.</b> Perbandingan Konfigurasi Pengujian .....	V-19

## DAFTAR TABEL

<b>Tabel II - 1.</b> Penelitian Terdahulu.....	II-20
<b>Tabel III - 1.</b> BIO Tagged Format .....	III-3
<b>Tabel III - 2.</b> Confusion Matrix (Multiclass) .....	II-14
<b>Tabel III - 3.</b> Rancangan Tabel Percobaan Hyperparameter .....	III-17
<b>Tabel III - 4.</b> Alat Bantu Penelitian.....	III-18
<b>Tabel III - 5.</b> Tabel Hasil Pengujian .....	III-19
<b>Tabel III - 6.</b> Work Breakdown Structure (WBS) .....	III-21
<b>Tabel IV - 1.</b> Kebutuhan Fungsional .....	IV-2
<b>Tabel IV - 2.</b> Kebutuhan Non-Fungsional .....	IV-2
<b>Tabel IV - 3.</b> Definisi Aktor.....	IV-4
<b>Tabel IV - 4.</b> Definisi Use Case.....	IV-4
<b>Tabel IV - 5.</b> Skenario Use Case Memuat Data.....	IV-5
<b>Tabel IV - 6.</b> Skenario Use Case Menampilkan Hasil Prediksi.....	IV-6
<b>Tabel IV - 7.</b> Rencana Pengujian Use Case Memuat Data .....	IV-17
<b>Tabel IV - 8.</b> Rencana Pengujian Use Case Menampilkan Hasil Prediksi Entitas .....	IV-17
<b>Tabel IV - 9.</b> Pengujian Use Case Memuat Data.....	IV-18
<b>Tabel IV - 10.</b> Pengujian Use Case Menampilkan Hasil Prediksi Entitas.....	IV-18
<b>Tabel V - 1.</b> Tabel Konfigurasi Percobaan.....	V-2
<b>Tabel V - 2.</b> Tabel Hasil Pengujian Konfigurasi I.....	V-5
<b>Tabel V - 3.</b> Tabel Hasil Pengujian Konfigurasi II .....	V-7
<b>Tabel V - 4.</b> Tabel Hasil Pengujian Konfigurasi III .....	V-9
<b>Tabel V - 5.</b> Tabel Hasil Pengujian Konfigurasi IV .....	V-11
<b>Tabel V - 6.</b> Tabel Hasil Pengujian Konfigurasi V .....	V-13
<b>Tabel V - 7.</b> Tabel Hasil Pengujian Konfigurasi VI.....	V-15
<b>Tabel V - 8.</b> Tabel Hasil Pengujian Konfigurasi VII .....	V-17
<b>Tabel V - 9.</b> Tabel Hasil Pengujian Konfigurasi VIII .....	V-19



# **BAB I**

## **PENDAHULUAN**

### **1.1 Pendahuluan**

Bab ini akan menguraikan elemen-elemen utama yang menjadi landasan penelitian ini. Pembahasan akan mencakup beberapa aspek kunci, di antaranya latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, dan manfaat penelitian. Uraian mengenai aspek-aspek fundamental tersebut akan berfungsi sebagai kerangka dasar yang mengarahkan keseluruhan proses penelitian ini.

### **1.2 Latar Belakang Masalah**

Dokumen tekstual sering kali mengandung beragam informasi, termasuk nama individu, tempat, atau lembaga. Proses manual untuk mengekstrak informasi dari teks yang panjang membutuhkan waktu dan tenaga, karena pembaca harus menelaah seluruh konten dengan seksama. Kompleksitas dan volume dokumen yang terus meningkat membuat pendekatan manual menjadi semakin tidak efisien. Sebagai solusi, teknologi *Named Entity Recognition* (NER) dikembangkan untuk mengotomatisasi proses identifikasi dan ekstraksi entitas penting seperti nama orang, tempat dan lembaga dari dokumen teks. Implementasi NER memungkinkan analisis dan pengambilan informasi yang lebih cepat dan efisien, secara signifikan mengurangi waktu yang dibutuhkan untuk memproses dan memahami konten dokumen yang besar.

Dalam konteks bahasa Indonesia, penerapan *Named Entity Recognition* (NER) menghadapi tantangan unik yang memerlukan perhatian khusus. Struktur linguistik

bahasa Indonesia yang kompleks, dengan variasi dialek dan pengaruh bahasa daerah, menciptakan kompleksitas tambahan dalam identifikasi entitas bernama (Alfina et al., 2017). Penggunaan model berbasis *Conditional Random Fields* (CRF) dengan fitur linguistik khusus bahasa Indonesia dapat meningkatkan akurasi NER, namun masih terbatas pada domain tertentu (Wibisono & Khodra, 2018). Lebih lanjut, integrasi teknik *word embedding* yang sesuai dengan model yang dibangun dapat secara signifikan meningkatkan kinerja sistem NER untuk bahasa Indonesia, terutama dalam menangani entitas yang jarang muncul atau ambigu (Susanty & Sukardi, 2021).

Meskipun demikian, perkembangan terkini dalam teknologi *deep learning* membuka peluang baru untuk meningkatkan efektivitas NER pada teks berbahasa Indonesia. Model bahasa pra-terlatih khusus bahasa Indonesia, seperti IndoBERT, memiliki potensi signifikan untuk meningkatkan kinerja NER dibandingkan dengan model multibahasa. Namun, implementasi IndoBERT untuk NER masih memerlukan penyesuaian dan peningkatan lebih lanjut, terutama dalam hal *fine-tuning* untuk tugas spesifik NER (Wilie et al., 2020).

Implementasi IndoBERT dalam konteks *Named Entity Recognition* (NER) untuk bahasa Indonesia menawarkan prospek signifikan dalam peningkatan presisi dan efisiensi. Meski demikian, tantangan tetap ada dalam hal generalisasi model terhadap berbagai domain dan dialek bahasa Indonesia. Pentingnya pengembangan dataset NER yang lebih komprehensif dan representatif untuk bahasa Indonesia guna mendukung evaluasi dan peningkatan model yang lebih akurat (Cahyawijaya et al., 2021).

Berdasarkan uraian di atas, dapat disimpulkan bahwa *Named Entity Recognition* (NER) untuk bahasa Indonesia masih memiliki ruang pengembangan yang signifikan, terutama dengan pemanfaatan model bahasa pra-terlatih seperti IndoBERT. Meskipun penelitian terdahulu telah menunjukkan potensi peningkatan kinerja melalui *fine-tuning*, masih terdapat kelemahan dalam hal generalisasi model terhadap variasi domain dalam Bahasa Indonesia, yang mencakup teks formal maupun informal seperti bahasa gaul atau slang. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan pendekatan *fine-tuning* yang optimal pada arsitektur IndoBERT untuk tugas NER, dengan fokus pada peningkatan kinerja, efisiensi, dan kemampuan adaptasi model terhadap variasi teks berbahasa Indonesia, dibandingkan dengan hasil penelitian sebelumnya yang menjadi sumber acuan.

### **1.3 Rumusan Masalah**

Berdasarkan latar belakang tersebut maka didapatlah rumusan masalah sebagai berikut:

1. Bagaimana kemampuan penggunaan model arsitektur IndoBERT dalam meningkatkan kinerja *Named Entity Recognition* (NER) pada teks berbahasa Indonesia?
2. Bagaimana strategi *fine-tuning* yang dilakukan agar dapat meningkatkan efisiensi model NER dalam menangani variasi teks berbahasa Indonesia yang kompleks?

#### 1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dijelaskan, maka terdapat tujuan:

1. Mengetahui kemampuan penggunaan model IndoBERT dalam meningkatkan kinerja *Named Entity Recognition* (NER) pada teks berbahasa Indonesia.
2. Mengetahui strategi *fine-tuning* yang dilakukan guna meningkatkan akurasi dan efisiensi model NER dalam menangani variasi teks berbahasa Indonesia yang kompleks.

#### 1.5 Manfaat Penelitian

1. Bagi peneliti, penelitian ini memberikan kontribusi pada pengembangan teknologi NER untuk bahasa Indonesia, khususnya dalam konteks penggunaan model bahasa pra-terlatih seperti IndoBERT.
2. Bagi peneliti selanjutnya, penelitian ini dapat digunakan sebagai referensi yang berharga dalam mengembangkan pendekatan *Named Entity Recognition* (NER) dengan metode dan arsitektur lainnya.
3. Hasil penelitian ini dapat digunakan untuk meningkatkan aplikasi NER dalam berbagai domain di Indonesia, seperti dalam bidang media, analisis data teks, dan otomatisasi layanan informasi. Ini akan membantu dalam pengembangan sistem yang lebih akurat dan efisien untuk pengenalan entitas di teks bahasa Indonesia.
4. Dengan meningkatnya kemampuan sistem NER untuk bahasa Indonesia, masyarakat Indonesia dapat memperoleh manfaat dari aplikasi teknologi yang

lebih canggih dalam berbagai aspek kehidupan, termasuk pemrosesan informasi dan komunikasi.

## **1.6 Batasan Masalah**

Adapun beberapa batasan masalah yang menjadi batas atau limitasi dalam penelitian ini adalah sebagai berikut:

1. Penelitian ini difokuskan pada penggunaan model IndoBERT *Base + Phase 2*, dan tidak mencakup analisis terhadap model bahasa lainnya.
2. Evaluasi dilakukan terhadap dataset NER yang terbatas pada teks berbahasa Indonesia dengan kategori entitas tertentu, seperti nama orang, lokasi dan organisasi.
3. Penelitian ini lebih berfokus pada pengembangan strategi *fine-tuning* menggunakan model asitektur IndoBERT.
4. Penilaian kinerja model akan dilakukan dengan menggunakan metrik standar seperti *accuracy*, *precision*, *recall*, dan *F1-Score*.

## **1.7 Sistematika Penulisan**

Sistematika penulisan tugas akhir mengikuti standar penulisan tugas akhir Fakultas Ilmu Komputer Universitas Sriwijaya yaitu sebagai berikut:

## **BAB I. PENDAHULUAN**

Bab ini akan mengulas mengenai latar belakang, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah, serta rencana sistematika penelitian yang akan menjadi fokus utama dalam penelitian ini.

## **BAB II. KAJIAN LITERATUR**

Bab ini membahas landasan teori yang digunakan dalam penelitian, seperti definisi metode dan arsitektur IndoBERT, *fine-tuning*, termasuk definisi *Named Entity Recognition* (NER), serta beberapa literatur yang relevan dengan penelitian ini.

## **BAB III. METODOLOGI PENELITIAN**

Bab ini akan menjelaskan rangkaian proses yang ditempuh selama penelitian, meliputi tahap pengumpulan data, analisis data, serta perancangan perangkat lunak. Setiap tahap akan diuraikan berdasarkan kerangka kerja yang telah direncanakan.

## **BAB IV. PENGEMBANGAN PERANGKAT LUNAK**

Bab ini akan membahas tentang proses perancangan perangkat lunak yang akan dibangun, mulai dari analisis kebutuhan hingga tahap pengujian untuk menilai hasil pengembangan perangkat lunak.

## **BAB V. HASIL DAN ANALISIS PENELITIAN**

Bab ini menyajikan hasil dan analisis dari penelitian yang telah dilaksanakan sesuai dengan langkah dan metode yang telah direncanakan sebelumnya. Analisis ini digunakan sebagai landasan untuk menarik kesimpulan dari penelitian ini.

## **BAB VI. KESIMPULAN DAN SARAN**

Bab ini membahas kesimpulan yang diambil berdasarkan uraian dalam bab sebelumnya serta saran yang diberikan berdasarkan penelitian yang telah dilakukan.

### **1.8 Kesimpulan**

Pada bab ini telah dijelaskan latar belakang, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah serta sistematika penelitian yang akan dijadikan sebagai pokok pikiran penelitian *Named Entity Recognition* (NER) pada teks berbahasa Indonesia dengan *fine-tuning* IndoBERT.

## DAFTAR PUSTAKA

- Alfina, I., Manurung, R., & Fanany, M. I. (2017). DBpedia entities expansion in automatically building dataset for Indonesian NER. *2016 International Conference on Advanced Computer Science and Information Systems, ICACISIS 2016*, 335–340. <https://doi.org/10.1109/ICACISIS.2016.7872784>
- Azarine, I. S., Bijaksana, M. A., & Asror, I. (2019). Named Entity Recognition For Indonesian Text Using Hidden Markov Model. *2019 7th International Conference on Information and Communication Technology (ICoICT)*, 1–5.
- Azzahra, N. S., Ibrohim, M. O., Fahmi, J., Apriyanto, B. F., & Riandi, O. (2020). Developing name entity recognition for structured and unstructured text formatting dataset. *2020 5th International Conference on Informatics and Computing, ICIC 2020*. <https://doi.org/10.1109/ICIC50835.2020.9288566>
- Bock, K., & Garnsey, S. M. (2008). Language Processing. *A Companion to Cognitive Science*, 226–234. <https://doi.org/10.1002/9781405164535.ch14>
- Budi, I., Bressan, S., Wahyudi, G., Hasibuan, Z. A., & Nazief, B. A. A. (2005). Named Entity Recognition for the Indonesian language: Combining contextual, morphological and part-of-speech features into a knowledge engineering approach. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3735 LNAI, 57–69. [https://doi.org/10.1007/11563983\\_7](https://doi.org/10.1007/11563983_7)
- Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M. L., Purwarianti, A., & Fung, P. (2021). IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 8875–8898.



<https://doi.org/10.18653/v1/2021.emnlp-main.699>

Chowdhary, K. R. (2020). Fundamentals of artificial intelligence. In *Fundamentals of Artificial Intelligence*. <https://doi.org/10.1007/978-81-322-3972-7>

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Naacl-Hlt 2019, Mlm*, 4171–4186. <https://aclanthology.org/N19-1423.pdf>

Franke, M., & Degen, J. (2023). The softmax function : Properties , motivation , and interpretation. *PsyArXiv*, 1–23.

Gunawan, W., Suhartono, D., Purnomo, F., & Ongko, A. (2018). Named-Entity Recognition for Indonesian Language using Bidirectional LSTM-CNNs. *Procedia Computer Science*, 135, 425–432. <https://doi.org/10.1016/j.procs.2018.08.193>

Hakimin, K., Jaroji, & Subandri, M. A. (2021). Penerapan Metode Rational Unified Process ( Rup ) Pada Pembuatan Aplikasi Public Speaking. *Seminar Nasional Industri Dan Teknologi (SNIT)*, 250–259. <https://snit-polbeng.org/eprosiding/index.php/snit/article/download/225/178>

Huang, J., Li, C., Subudhi, K., Jose, D., Balakrishnan, S., Chen, W., Peng, B., Gao, J., & Han, J. (2020). *Few-Shot Named Entity Recognition: A Comprehensive Study*. <http://arxiv.org/abs/2012.14978>

Hugging Face. (2023). *No Title*. NLP Course - WordPiece Tokenization. <https://huggingface.co/learn/nlp-course/chapter6/6?fw=pt#wordpiece-tokenization>

Katiyar, A., & Cardie, C. (2018). Nested named entity recognition revisited. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of*

*the Conference*, 1, 861–871. <https://doi.org/10.18653/v1/n18-1079>

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, 757–770.

<https://doi.org/10.18653/v1/2020.coling-main.66>

Krstinić, D., Braović, M., Šerić, L., & Božić-Štulić, D. (2020). *Multi-label Classifier Performance Evaluation with Confusion Matrix*. 01–14.

<https://doi.org/10.5121/csit.2020.100801>

Kuriakose, J. (2019). *No Title*. BIO / IOB Tagged Text to Original Text.

<https://medium.com/analytics-vidhya/bio-tagged-text-to-original-text-99b05da6664>

Li, J., Sun, A., Han, J., & Li, C. (2022). A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1), 50–70. <https://doi.org/10.1109/TKDE.2020.2981314>

Mohiuddin, K., Welke, P., Alam, M. A., Martin, M., Alam, M. M., Lehmann, J., & Vahdati, S. (2023). Retention Is All You Need. *International Conference on Information and Knowledge Management, Proceedings, Nips*, 4752–4758.

<https://doi.org/10.1145/3583780.3615497>

Muller, B., Anastasopoulos, A., Sagot, B., & Seddah, D. (2021). When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 448–462.

<https://doi.org/10.18653/v1/2021.naacl-main.38>

- Mulyani, A., Septiana, Y., & Alamsyah, R. (2022). Rancang Bangun Sistem Informasi Pengaduan Layanan Menggunakan Metode Rational Unified Process. *Jurnal Algoritma*, 19(2), 722–728. <https://doi.org/10.33364/algoritma/v.19-2.1199>
- Patil, N., Patil, A., & Pawar, B. V. (2020). Named Entity Recognition using Conditional Random Fields. *Procedia Computer Science*, 167(2019), 1181–1188. <https://doi.org/10.1016/j.procs.2020.03.431>
- Rahmayadi, E., & Sujadi, H. (2022). Pengembangan Sistem Informasi Manajemen Pengadaan Pada PT Wijaya Karya Beton, Tbk., PPB Majalengka. *Journal of Information System and Technology*, 1(1), 11–17. <https://doi.org/10.56916/jistec.v1i1.81>
- Ramli, N. E., Yahya, Z. R., & Said, N. A. (2022). Confusion Matrix as Performance Measure for Corner Detectors. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 29(1), 256–265. <https://doi.org/10.37934/araset.29.1.256265>
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8, 842–866. [https://doi.org/10.1162/tacl\\_a\\_00349](https://doi.org/10.1162/tacl_a_00349)
- Sasaki, Y. (2007). The truth of the F-measure. *Teach Tutor Mater*, 1–5. <http://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf>
- Susanty, M., & Sukardi, S. (2021). Perbandingan Pre-trained Word Embedding dan Embedding Layer untuk Named-Entity Recognition Bahasa Indonesia. *Petir*, 14(2), 247–257. <https://doi.org/10.33322/petir.v14i2.1164>
- Tia, T., Nuryasin, I., & Maskur, M. (2020). Model Simulasi Rational Unified Process

(RUP) Pada Pengembangan Perangkat Lunak. *Jurnal Repositor*, 2(4), 485–494.  
<https://doi.org/10.22219/repositor.v2i4.390>

Wiawan, M. I., Fadillah, R., & Yusup, D. M. (2023). Sistem Informasi Pelayanan Administrasi Berbasis Website Dengan Metode Rup (Rational Unified Process) (Studi Kasus Kantor Urusan Agama Kecamatan Panjalu). *EDUSAINTEK: Jurnal Pendidikan, Sains Dan Teknologi*, 10(3), 758–776.  
<https://doi.org/10.47668/edusaintek.v10i3.865>

Wibisono, Y., & Khodra, M. L. (2018). Pengenalan Entitas Bernama Otomatis untuk Bahasa Indonesia dengan Pendekatan Pembelajaran Mesin. *Seminar Tahunan Linguistik 2018*, 1–5.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). *IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding*. 843–857. <http://arxiv.org/abs/2009.05387>

Yadav, V., & Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*, 2145–2158.

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, 1168(2). <https://doi.org/10.1088/1742-6596/1168/2/022022>