

**IMPLEMENTASI NAMED ENTITY RECOGNITION UNTUK
IDENTIFIKASI DIAGNOSIS MEDIS MENGGUNAKAN
MODEL MULTILINGUAL BERT**



SKRIPSI

**STEVE CHRISTIAN Y. P.
00000058797**

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2025**

**IMPLEMENTASI NAMED ENTITY RECOGNITION UNTUK
IDENTIFIKASI DIAGNOSIS MEDIS MENGGUNAKAN
MODEL MULTILINGUAL BERT**



SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh
Gelar Sarjana Komputer (S.Kom.)

**STEVE CHRISTIAN Y. P.
00000058797**

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2025**

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Steve Christian Y. P.

Nomor Induk Mahasiswa : 00000058797

Program Studi : Informatika

Skripsi dengan judul:

Implementasi Named Entity Recognition untuk Identifikasi Diagnosis Medis pada Teks Berbahasa Indonesia Menggunakan Multilingual BERT Berbasis BIO Tagging dan FHIR

merupakan hasil karya saya sendiri bukan plagiat dari laporan karya tulis ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan maupun dalam penulisan laporan karya tulis ilmiah, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah yang telah saya tempuh.

Tangerang, 4 Juli 2025



(Steve Christian Y. P.)

HALAMAN PENGESAHAN

Skripsi dengan judul

IMPLEMENTASI NAMED ENTITY RECOGNITION UNTUK IDENTIFIKASI DIAGNOSIS MEDIS MENGGUNAKAN MODEL MULTILINGUAL BERT

oleh

Nama : Steve Christian Y. P.
NIM : 00000058797
Program Studi : Informatika
Fakultas : Fakultas Teknik dan Informatika

Telah diujikan pada hari Senin, 14 Juli 2025

Pukul 08.00 s/d 10.00 dan dinyatakan

LULUS

Dengan susunan penguji sebagai berikut

Ketua Sidang

(Januar Wahjudi, S.Kom., M.Sc.)

NIDN: 0330017201

Penguji

(Sy Yuliani Yakub, S.Kom., M.T. PhD)

NIDN: 0411037904

Pembimbing

(Dr. Ivransa Zuhdi Pane, B. Eng., M. Eng.)

NIDN: 8812520016

Ketua Program Studi Informatika,

(Arya Wicaksana, S.Kom., M.Eng.Sc., OCA)

NIDN: 0315109103

**HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK
KEPENTINGAN AKADEMIS**

Yang bertanda tangan di bawah ini:

Nama : Steve Christian Y. P.
NIM : 00000058797
Program Studi : Informatika
Jenjang : S1
Judul Karya Ilmiah : Implementasi Named Entity
Recognition untuk Identifikasi
Diagnosis Medis pada Teks Berbahasa
Indonesia Menggunakan Multilingual
BERT Berbasis BIO Tagging dan
FHIR

Menyatakan dengan sesungguhnya bahwa saya bersedia (**pilih salah satu**):

- ☒ Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.
- ☐ Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) **.
- ☐ Lainnya, pilih salah satu:
- Hanya dapat diakses secara internal Universitas Multimedia Nusantara
 - Embargo publikasi karya ilmiah dalam kurun waktu tiga tahun.

Tangerang, 4 Juli 2025

Yang menyatakan



Steve Christian Y. P.

HALAMAN PERSEMBAHAN / MOTTO

"A good name is to be more desired than great wealth, Favor is better than silver and gold."

Proverbs 22:1 (NASB)



KATA PENGANTAR

Puji Syukur atas berkat dan rahmat kepada Tuhan Yang Maha Esa, atas selesainya penulisan laporan Skripsi ini dengan judul: Implementasi Named Entity Recognition Untuk Identifikasi Diagnosis Medis menggunakan Model Multilingual BERT dilakukan untuk memenuhi salah satu syarat untuk mencapai gelar Sarjana Komputer Jurusan Informatika Pada Fakultas Teknik dan Informatika Universitas Multimedia Nusantara. Saya menyadari bahwa, tanpa bantuan dan bimbingan dari berbagai pihak, dari masa perkuliahan sampai pada penyusunan laporan Skripsi ini, sangatlah sulit bagi saya untuk menyelesaikan laporan Skripsi ini. Oleh karena itu, saya mengucapkan terima kasih kepada:

Mengucapkan terima kasih

1. Bapak Dr. Ir. Andrey Andoko, M.Sc., selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Bapak Arya Wicaksana, S.Kom., M.Eng.Sc., OCA, selaku Ketua Program Studi Informatika Universitas Multimedia Nusantara.
4. Bapak Dr. Ivansa Zuhdi Pane, B.Eng., M.Eng., sebagai Pembimbing pertama yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Keluarga saya yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan tugas akhir ini.

Semoga karya ilmiah ini dapat memberikan manfaat dan kontribusi positif, baik bagi penulis sendiri, dunia pendidikan, maupun pihak-pihak lain yang membutuhkan informasi terkait. Penulis juga berharap laporan ini dapat menjadi referensi serta inspirasi untuk penelitian selanjutnya dalam bidang yang relevan.

Tangerang, 4 Juli 2025

Steve Christian Y. P.

IMPLEMENTASI NAMED ENTITY RECOGNITION UNTUK IDENTIFIKASI DIAGNOSIS MEDIS MENGGUNAKAN MODEL MULTILINGUAL BERT

Steve Christian Y. P.

ABSTRAK

Penelitian ini mengembangkan sebuah *pipeline* otomatis berbasis Multilingual BERT dan skema BIO tagging untuk ekstraksi dan identifikasi diagnosis medis dari teks berbahasa Indonesia. Proses *pipeline* mencakup *preprocessing* data, pelabelan BIO, pelatihan model NER menggunakan mBERT, pemetaan ke kode SNOMED CT, dan konversi ke *resource Fast Healthcare Interoperability Resources* (FHIR). *Dataset* terdiri dari 9.000 kalimat yang mengandung diagnosis, dengan distribusi label yang beragam. Pengujian pada data nyata dari *website* kesehatan seperti *Halodoc* dan *Alodokter* menunjukkan model mampu mencapai akurasi sebesar 93.1%, precision 91.8%, recall 92.3%, dan F1-score 92% untuk label *Condition*. *Pipeline* juga diimplementasikan sebagai layanan *API* berbasis *Flask*. Tantangan utama meliputi keterbatasan kamus SNOMED CT lokal dan variasi istilah diagnosis. *Pipeline* ini berpotensi mendukung digitalisasi dan interoperabilitas data klinis di Indonesia.

Kata kunci: *BIO tagging, Fast Healthcare Interoperability Resources, Multilingual BERT, NER, SNOMED CT*



IMPLEMENTATION OF NAMED ENTITY RECOGNITION FOR MEDICAL DIAGNOSIS IDENTIFICATION USING MULTILINGUAL BERT MODEL

Steve Christian Y. P.

ABSTRACT

This study develops an automatic pipeline based on Multilingual BERT and the BIO tagging scheme for extracting and identifying medical diagnoses from Indonesian-language texts. The pipeline covers data preprocessing, BIO labeling, NER model training using mBERT, mapping to SNOMED CT codes, and conversion to FHIR resources. The dataset consists of 9,000 diagnosis sentences with diverse label distribution. Testing on real-world data from health websites such as Halodoc and Alodokter shows that the model achieves 93.1% accuracy, 91.8% precision, 92.3% recall, and 92% F1-score for the Condition label. The pipeline is also implemented as an API service using Flask. Major challenges include the limited availability of a local SNOMED CT dictionary and the wide variety of diagnosis terms. This pipeline has the potential to support the digitalization and interoperability of clinical data in Indonesia.

Keywords: *BIO tagging, Fast Healthcare Interoperability Resources, Multilingual BERT, NER, SNOMED CT*



DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	iv
HALAMAN PERSEMBAHAN/MOTO	v
KATA PENGANTAR	vi
ABSTRAK	vii
ABSTRACT	viii
DAFTAR ISI	ix
DAFTAR TABEL	x
DAFTAR GAMBAR	xi
DAFTAR KODE	xii
DAFTAR RUMUS	xiii
DAFTAR LAMPIRAN	xiii
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah	2
1.3 Batasan Permasalahan	2
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	3
1.6 Sistematika Penulisan	4
BAB 2 LANDASAN TEORI	5
2.1 Natural Language Processing (NLP)	5
2.2 Named Entity Recognition (NER)	5
2.3 Multilingual Bert (mBERT)	6
2.4 BIO Tagging	6
2.5 SNOMED CT	7
2.6 Fast Healthcare Interoperability Resources (FHIR)	7
2.7 Penelitian Terkait	8
BAB 3 METODOLOGI PENELITIAN	10
3.1 Sumber Data	11
3.2 Preprocessing Data	12
3.2.1 Isi Data Diagnosis dan Contoh Format BIO	12
3.3 Proses Named Entity Recognition (NER)	13
3.4 Pengembangan Model NER Berbasis Multilingual BERT	14
3.5 Pemetaan ke Kode SNOMED CT	16
3.6 Konversi Diagnosis ke Format FHIR Condition	16
3.6.1 Evaluasi Pipeline dan Analisis Akurasi	18
BAB 4 HASIL DAN DISKUSI	20
4.1 Hasil Perancangan	20
4.1.1 Hasil Implementasi Backend	20
4.2 Pembahasan Hasil Rancangan Sistem Ekstraksi Condition Berbasis NER BERT	22
4.3 Pengujian Sistem	32
4.4 Tampilan <i>Output</i> Sistem	36
4.4.1 Token-Label Diagnosis	36
4.4.2 Diagnosis <i>Entity</i> dan <i>Mapping</i> SNOMED CT	37
4.4.3 Output FHIR Condition	37
4.4.4 Kesimpulan Output Sistem	38
BAB 5 SIMPULAN DAN SARAN	39
5.1 Simpulan	39
5.2 Saran	40
DAFTAR PUSTAKA	41

DAFTAR TABEL

Tabel 2.1	Perbandingan Penelitian Terkait dengan Penelitian Ini	9
Tabel 3.1	Sumber Dataset Diagnosis Medis	11
Tabel 3.2	Contoh Kalimat Diagnosis dan Label BIO Tagging	13
Tabel 3.3	Contoh Pemetaan Diagnosis ke Kode SNOMED CT	16
Tabel 4.1	Confusion Matrix untuk Data Training	33
Tabel 4.2	Laporan Klasifikasi untuk Data <i>Training</i>	33
Tabel 4.3	Confusion Matrix untuk Data Validasi/Testing	33
Tabel 4.4	Laporan Klasifikasi untuk Data Validasi/Testing	34
Tabel 4.5	Confusion Matrix Token-level pada Dataset <i>Website</i> Kesehatan (Halodoc, Alodokter, Satu Sehat, dsb.)	34
Tabel 4.6	Laporan Klasifikasi pada Dataset <i>Website</i> Kesehatan	35
Tabel 4.7	Ringkasan Hasil Evaluasi <i>Pipeline</i> pada Berbagai <i>Dataset</i>	35



DAFTAR GAMBAR

Gambar 3.1	Diagram Alur Penelitian	10
Gambar 3.2	Diagram Alur Proses Named Entity Recognition (NER)	14
Gambar 3.3	<i>activity diagram</i> konversi diagnosis ke FHIR CONDITION	18
Gambar 4.1	Contoh pengujian <i>endpoint</i> API NER menggunakan metode POST dengan format JSON.	20
Gambar 4.2	Contoh hasil pengujian <i>endpoint</i> API NER menggunakan <i>Postman</i> . Permintaan (request) berisi diagnosis dalam format JSON, dan respons menampilkan hasil pelabelan token beserta label BIO (B- CONDITION, I-CONDITION, O) untuk setiap kata.	21
Gambar 4.3	Contoh <i>function</i> untuk <i>load dataset</i> berformat CoNLL	23
Gambar 4.4	Contoh <i>function</i> untuk <i>preprocessing</i> data yang akan masuk ke model BERT	24
Gambar 4.5	Contoh <i>function</i> untuk <i>labelling</i> dan tokenisasi data yang akan masuk ke model BERT	25
Gambar 4.6	Contoh <i>function</i> untuk <i>handle</i> NER sesuai model BERT	26
Gambar 4.7	Contoh <i>function</i> untuk metrik evaluasi dan inisialisasi model	26
Gambar 4.8	Contoh <i>function</i> untuk menentukan <i>hyperparameter</i> optimal untuk model <i>training</i> BERT	27
Gambar 4.9	Contoh <i>function</i> pelatihan model <i>Named Entity Recognition</i> dengan skema <i>K-Fold cross-validation</i> sebanyak <i>5fold</i>	28
Gambar 4.10	<i>Learning curve model</i> pada fold 1, menunjukkan peningkatan <i>F1-score</i> , akurasi, dan penurunan <i>validation loss</i> pada setiap <i>epoch</i>	29
Gambar 4.11	<i>Learning curve model</i> pada fold 2, menunjukkan tren performa yang stabil dan <i>robust</i> sepanjang pelatihan.	30
Gambar 4.12	<i>Learning curve model</i> pada fold 3, menunjukkan peningkatan <i>F1-score</i> dan akurasi secara bertahap di setiap <i>epoch</i>	30
Gambar 4.13	<i>Learning curve model</i> pada fold 4, dengan tren performa yang konsisten dan <i>robust</i>	31
Gambar 4.14	<i>Learning curve model</i> pada fold 5, memperlihatkan peningkatan performa dan kestabilan <i>pipeline</i> pada seluruh proses pelatihan.	32
Gambar 4.15	Contoh Output Token-Label Diagnosis	36
Gambar 4.16	Contoh Diagnosis <i>Entity</i> dan Hasil <i>Mapping</i> SNOMED CT	37
Gambar 4.17	Contoh <i>Output</i> FHIR CONDITION dalam Format JSON	38

UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR KODE



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR LAMPIRAN

Lampiran 1	Hasil Persentase Turnitin (Wajib Ada)	43
Lampiran 2	Formulir Bimbingan (Wajib Ada)	45
Lampiran 3	Transkrip Wawancara (optional)	47
Lampiran 4	Judul yang Panjang Melampaui Satu Baris Judul yang Panjang Melampaui Satu Baris	48

