

***CLINICAL NAMED ENTITY RECOGNITION* PADA
DATA BIOMEDIS MENGGUNAKAN VARIAN MODEL
BERT UNTUK KASUS BIOMEDIS**

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat

Memperoleh Gelar Sarjana Komputer



OLEH :

TIARA OKTARINA

09011182126028

**SISTEM KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA
TAHUN 2025**

HALAMAN PENGESAHAN

***CLINICAL NAMED ENTITY RECOGNITION* PADA DATA BIOMEDIS MENGGUNAKAN VARIAN MODEL BERT UNTUK KASUS BIOMEDIS**

Sebagai Salah Satu Syarat untuk penyelesaian studi di
Program Studi S1 Sistem Komputer

Oleh :

TIARA OKTARINA

09011182126028

Pembimbing

: Dr. Firdaus, S.T., M.Kom.

NIP. 197801212008121003

Mengetahui,

Ketua Jurusan Sistem Komputer



Dr. Ir. Sukemi, M.T.
196612032006041001

AUTHENTICATION PAGE

***CLINICAL NAMED ENTITY RECOGNITION ON
BIOMEDICAL DATA USING BERT MODEL VARIANTS
FOR BIOMEDICAL CASES***

*As one of the requirements for completing studies
in the Bachelor's Degree of Computer Systems Study Program*

By :

TIARA OKTARINA

09011182126028

***Supervisor* : Dr. Firdaus, S.T., M.Kom.
NIP. 197801212008121003**

***Approved by,
Head of Computer System Department***



**Dr. Ir. Sukemi, M.T.
196612032006041001**

HALAMAN PERSETUJUAN

Telah diuji dan lulus pada :

Hari : Kamis ✓
Tanggal : 26 Juni 2025 ✓
Tim penguji

1. Ketua sidang : Dr. Rossi Passarella, M. Eng.
2. Penguji sidang : Sutarno, M. T.
3. Pembimbing : Dr. Firdaus, S.T., M.Kom



A.

Firdaus

Mengetahui, 28/7/25
Ketua Jurusan Sistem Komputer


Dr. Ir. H. Sukemi, M.T.
NIP. 196612032006041001

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Tiara Oktarina
Nim : 09011182126028
Judul tugas akhir : *Clinical Named Entity Recognition* Pada Data Biomedis
Menggunakan Varian Model BERT untuk kasus biomedis

Hasil pemeriksaan iThenticate/Turnitin: 2 %

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

Demikian, pernyataan ini dibuat dalam keadaan sadar dan tidak dipaksakan.



Indralaya, Juli 2025

Tiara Oktarina
NIM. 09011182126028

HALAMAN PERSEMBAHAN

Dengan segala kerendahan hati dan rasa syukur yang mendalam kepada Allah SWT, tugas akhir ini kupersembahkan untuk orang-orang tercinta dalam hidupku :

1. Untuk Orang Tuaku Tercinta, Terutama Mama.
Terima kasih atas kasih sayangmu, perjuanganmu, serta doamu yang selalu membimbingku dalam setiap langkah, dalam setiap pengambilan keputusan, terima kasih atas semua pengorbananmu yang bahkan tak mampu aku balas. Semoga skripsi ini menjadi salah satu bukti kecil atas keberhasilanmu dalam mendidikku, menjadi bukti bahwa perjuanganmu tidak sia-sia. Ma, bumi ini luas, aku berteduh di bawah doamu. Love u ma pak.
2. Untuk Nenekku Tersayang
Terima kasih banyak atas doa-doa tulus yang menenangkan hati, atas doamu pula langkahku dipermudah oleh Allah, Terima kasih nek atas semua kasih sayangmu membuatku selalu merasa aman dan dicintai, sudah merawatku sejak kecil. Skripsi ini aku persembahkan pula untukmu, orang tersayangku. Panjang umur yaa.
3. Untuk Adik dan Aa
Aku merasa bersyukur telah lahir dan memiliki adik dan aa di hidupku, terima kasih telah menemani hari-hariku, adik menjadi salah satu orang yang membuat aku tidak menyerah dalam proses penyelesaian skripsi ini, Aa menjadi orang yang memotivasiku untuk menjadi lebih baik setiap harinya.
4. Untuk Keluarga Besar
Terima kasih atas semua doa dan dukungan selama ini, skripsi ini pula aku persembahkan untuk keluarga besar, hal kecil yang semoga membuat kalian bangga.

MOTTO

“Allah tidak akan membebani seseorang melainkan sesuai dengan kesanggupannya”
(Q.S Al-Baqarah 2:286)

“Pada akhirnya, ini semua hanyalah permulaan” (Nadin Amizah)

“Long Story Short, I Survived” (Taylor Swift)

“You’re always one decision away from a completely different life”
(Maudy Ayunda)

KATA PENGANTAR

Assalamu'alaikum Wr.Wb.

Puji dan syukur atas kehadiran Allah SWT yang telah melimpahkan rahmat, kasih sayang dan karunia-Nya, sehingga penulis dapat menyelesaikan Tugas Akhir dengan judul “*Clinical Named Entity Recognition* Pada Data Biomedis Menggunakan Varian Model Bert Untuk Kasus Biomedis”.

Dalam laporan ini, penulis menjelaskan pemahaman tentang pemodelan *Named Entity Recognition* (NER) pada data biomedis menggunakan model *BioBERT*, *Clinical BERT*, dan *BlueBERT* serta solusi untuk meningkatkan pengelolaan analisis data kesehatan. Penulis berharap agar tulisan ini dapat bermanfaat bagi banyak orang.

Selama penulisan tugas akhir ini, penulis banyak mendapatkan doa, bantuan, serta saran dari semua pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan banyak terima kasih kepada :

1. Allah SWT yang telah melimpahkan berkat serta nikmat kesehatan dan kesempatan kepada penulis untuk menyelesaikan tugas akhir.
2. Orang tua, saudara dan seluruh keluarga besar yang telah mendoakan, memberikan motivasi, dan mendukung penulis.
3. Bapak Prof. Dr. Erwin, S.Si., M.Si. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya dan selaku Dosen Pembimbing Akademik.
4. Bapak Dr. Ir. Sukemi, M, T., selaku Ketua Jurusan Sistem Komputer Fakultas Ilmu Komputer Universitas Sriwijaya.
5. Bapak Dr. Firdaus, S.T., M.Kom. sebagai Dosen Pembimbing Tugas Akhir yang telah meluangkan waktu untuk membimbing, memberikan saran, motivasi, dan bimbingan terbaik kepada penulis dalam menyelesaikan Tugas Akhir ini.
6. Ibu Prof. Dr. Ir. Siti Nurmaini, M. T., dan Bapak Prof. Dr. Ir. Bambang Tutuko, M. T selaku dosen IsysRG.

7. Ibu Anggun Islami, M.Kom selaku mentor divisi data tabular yang telah memberikan bantuan dan masukan dalam penelitian skripsi ini.
8. Ibu Dr. Ade Iriani Sapitri, M.Kom, Bapak M. Naufal Rahmatullah, Ibu Akhiar Wista Arum, M.Kom M.T, dan Ibu Annisa darmawahyuni, M.Kom selaku mentor yang telah memberikan bantuan dan masukan dalam penelitian skripsi ini.
9. Staf administrasi Fakultas Ilmu Komputer Universitas Sriwijaya Jurusan Sistem Komputer Kampus Indralaya yang telah memberikan kemudahan dalam hal administrasi sehingga penulis dapat membuat proposal tugas akhir ini dengan lancar.
10. Teman-teman IsysRG terutama Divisi *Text Processing*, Tria Lailani, M. Azriel Apriadi, Keisyah Syabinatullah, dan Indah Gala Putri.
11. Teman seperjuangan Sania Fatimah Azzahra, Zalfa Amira, dan Ferdi Aprian.
12. Teman-teman Sistem Komputer 2021 Indralaya
13. Seluruh pihak yang membantu penulis selama proses pembuatan proposal tugas akhir yang tidak dapat disebutkan satu persatu.

Penulis menyadari bahwa laporan ini masih sangat jauh dari kata sempurna. Oleh karena itu, penulis menyadari adanya banyak kekurangan dan kesalahan dalam penulisan laporan ini. Untuk itu kritik dan saran yang membangun sangat diharapkan penulis untuk perbaikan laporan-laporan di masa yang akan datang. Akhir kata penulis berharap, semoga laporan ini dapat bermanfaat bagi semua orang.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Indralaya, Juli 2025

Penulis,

Tiara Oktarina

NIM. 09011182126028

CLINICAL NAMED ENTITY RECOGNITION PADA DATA BIOMEDIS MENGGUNAKAN VARIAN MODEL BERT UNTUK KASUS BIOMEDIS

Tiara Oktarina (09011182126028)

Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya

Email : tiara.oktarina2003@gmail.com

ABSTRAK

Penelitian ini bertujuan untuk mengembangkan sistem yang mampu mengidentifikasi dan mengklasifikasikan entitas medis dari teks biomedis tidak terstruktur, guna mendukung analisis klinis dan riset kesehatan. penulis melatih dan mengevaluasi tiga model berbasis BERT, yaitu *BioBERT*, *Clinical BERT*, dan *BlueBERT*, dalam tugas *Clinical Named Entity Recognition* (CNER). Performa model diukur menggunakan metrik *precision*, *recall*, dan *F1-Score* pada tiga *dataset* biomedis: NCBI, BC2GM, dan JNLPBA. Hasil menunjukkan *BioBERT* secara konsisten memberikan kinerja terbaik di sebagian besar *dataset*. Pada NCBI, *BioBERT* mencapai *F1-Score* tertinggi 93%, mengungguli *Clinical BERT* dan *BlueBERT* (91%). Di BC2GM, *BioBERT* juga unggul dengan 91%, sementara model lain mencapai 90%. *Dataset* JNLPBA terbukti lebih menantang, dengan *F1-Score* tertinggi 80% (*BioBERT*). Peningkatan performa signifikan, hingga 10% pada beberapa kasus, terlihat pada percobaan akhir berkat penerapan *fine-tuning* penuh. Penelitian ini berkontribusi dalam mengidentifikasi model *transformer* yang paling optimal untuk aplikasi ekstraksi informasi otomatis di bidang kesehatan.

Kata Kunci : *Clinical Named Entity Recognition, Dataset Biomedis, Transformer, BioBERT, Clinical BERT, BlueBERT*

CLINICAL NAMED ENTITY RECOGNITION ON BIOMEDICAL DATA USING BERT MODEL VARIANTS FOR BIOMEDICAL CASES

Tiara Oktarina (09011182126028)

Computer System Department, Computer Science Faculty, Sriwijaya University

Email : tiara.oktarina2003@gmail.com

ABSTRACT

This study aims to develop a system capable of identifying and classifying medical entities from unstructured biomedical texts, thereby supporting Clinical analysis and health research. The author trained and evaluated three BERT-based models: BioBERT, Clinical BERT, and BlueBERT, for the task of Clinical Named Entity Recognition (CNER). Model performance was measured using precision, recall, and F1-Score metrics on three biomedical datasets: NCBI, BC2GM, and JNLPBA. The results consistently show that BioBERT delivers the best performance across most datasets. On NCBI, BioBERT achieved the highest F1-Score of 93%, outperforming Clinical BERT and BlueBERT (both 91%). In BC2GM, BioBERT also excelled with 91%, while other models reached 90%. The JNLPBA dataset proved more challenging, with BioBERT achieving the highest F1-Score of only 80%. A significant performance improvement, up to 10% in some cases, was observed in the final experiments due to the application of full fine-tuning. This research contributes to identifying the most optimal transformer model for automated information extraction applications in healthcare.

Keywords : *Clinical Named Entity Recognition, Dataset Biomedic, Transformer, BioBERT, Clinical BERT, BlueBERT*

DAFTAR ISI

HALAMAN PERSETUJUAN.....	Kesalahan! Bookmark tidak ditentukan.
HALAMAN PERNYATAAN.....	Kesalahan! Bookmark tidak ditentukan.
HALAMAN PERSEMBAHAN.....	v
KATA PENGANTAR.....	vi
ABSTRAK.....	viii
ABSTRACT	ix
DAFTAR ISI.....	x
DAFTAR GAMBAR.....	xiv
DAFTAR TABEL.....	xvi
BAB I.....	1
PENDAHULUAN.....	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	2
1.3. Batasan Masalah.....	3
1.4. Tujuan Penelitian	3
1.5. Metodologi Penulisan.....	4
1.5.1. Metode Studi Pustaka dan Literatur	4
1.5.2. Metode Konsultasi	4
1.5.3. Metode Pembuatan Model	4
1.5.4. Metode Pengujian dan Validasi	4
1.5.5. Metode Hasil dan Analisa	4
1.5.6. Metode Penarikan Kesimpulan dan Saran	5
1.6. Sistematika Penulisan.....	5
BAB II	6
TINJAUAN PUSTAKA	6
2.1. Penelitian Terdahulu	6

2.2.	<i>Text Preprocessing</i>	7
2.3.	<i>Text Normalization</i>	8
2.4.	<i>Text Tokenization</i>	8
2.5.	<i>Padding</i>	8
2.6.	<i>Word Embeddings</i>	9
2.7.	<i>Modelling</i>	9
2.8.	<i>Deep Learning</i>	9
2.9.	<i>Data Terstruktur</i>	10
2.10.	<i>Named Entity Recognition</i>	10
2.11.	<i>Natural Language Processing</i>	10
2.12.	<i>Clinical Named Entity Recognition</i>	11
2.13.	<i>Bidirectional Encoder Representations from Transformer</i>	11
2.14.	<i>BioBERT</i>	11
2.15.	<i>Clinical BERT</i>	12
2.16.	<i>BlueBERT</i>	12
2.17.	<i>Transformer</i>	12
2.18.	<i>Dataset Joint Workshop on Natural Language Processing in Biomedicine and its Applications</i>	13
2.19.	<i>Dataset National Center For Biotechnology Information</i>	13
2.20.	<i>Dataset BioCreative II Gene Mention</i>	13
2.21.	<i>Evaluasi Model</i>	14
2.22.	<i>Precision</i>	14
2.23.	<i>Recall</i>	14
2.24.	<i>F1-Score</i>	15
BAB III		16
METODOLOGI PENELITIAN		16
3.1.	<i>Kerangka Kerja</i>	16
3.2.	<i>Akuisisi Data</i>	17
3.3.	<i>Data Splitting</i>	18
3.4.	<i>Exploratory Data Analysis</i>	18
3.4.1.	<i>Descriptive Analysys</i>	19
3.4.2.	<i>Visualization Of Word Distribution</i>	20
3.4.3.	<i>Wordcloud Visualization Of Dominan Words</i>	21

3.4.4.	<i>Visualization Of Entity Label Percentage</i>	22
3.5.	<i>Data Preprocessing</i>	24
3.5.1.	<i>Text Normalization</i>	24
3.5.2.	<i>Tokenization</i>	25
3.5.3.	<i>Word Indexing</i>	26
3.5.4.	<i>Padding</i>	27
3.6.	<i>Modelling</i>	27
3.6.1.	Model <i>BioBERT</i>	28
3.6.2.	Model <i>Clinical BERT</i>	28
3.7.	Penambahan Model <i>BlueBERT</i> Sebagai Komparasi.....	29
3.8.	Evaluasi Model.....	30
BAB IV		31
HASIL DAN ANALISIS		31
4.1.	Skenario Percobaan.....	31
4.2.	Hasil Pengujian pada <i>Dataset NCBI</i>	31
4.3.	Hasil Pengujian pada <i>Dataset BC2GM</i>	37
4.4.	Hasil Pengujian pada <i>Dataset JNLPBA</i>	42
4.5.	Grafik Pada Data <i>Training</i> dan <i>Validation</i>	49
4.6.	<i>Macro average Dataset NCBI, BC2GM, dan JNLPBA</i>	52
4.6.1.	<i>Macro average</i> Ketiga <i>Dataset Biomedis</i> dengan <i>Batch Size 16</i>	53
4.6.2.	<i>Macro average</i> Ketiga <i>Dataset Biomedis</i> dengan <i>Batch Size 32</i>	60
4.7.	<i>Confusion matrix</i> pada Data <i>Testing</i>	66
4.7.1.	<i>Confusion matrix</i> pada <i>Dataset NCBI</i>	66
4.7.2.	<i>Confusion matrix</i> pada <i>Dataset BC2GM</i>	68
4.7.3.	<i>Confusion matrix</i> pada <i>Dataset JNLPBA</i>	71
4.8.	Durasi Waktu Training Model <i>BioBERT</i> dan <i>Clinical BERT</i>	73
4.9.	Peningkatan Metrik Evaluasi Model Pada 3 <i>Dataset Biomedis</i>	75
4.9.1.	Skenario Percobaan Akhir.....	75
4.9.2.	Hasil Pengujian pada <i>Dataset NCBI</i>	76
4.9.3.	Hasil Percobaan pada <i>Dataset BC2GM</i>	80
4.9.4.	Hasil Percobaan Pada <i>Dataset JNLPBA</i>	83
4.9.5.	<i>Macro average Dataset NCBI, BC2GM, dan JNLPBA</i>	90

4.9.6.	<i>Confusion matrix</i> pada Data <i>Testing</i> NCBI, BC2GM, Dan JNLPBA	97
4.9.7.	<i>Confusion matrix</i> pada <i>Dataset</i> NCBI	97
4.9.8.	<i>Confusion matrix</i> pada <i>Dataset</i> BC2GM	99
4.9.9.	<i>Confusion matrix</i> pada <i>Dataset</i> JNLPBA	101
4.9.10.	Durasi Waktu Model <i>Training BioBERT</i> , <i>Clinical BERT</i> , Dan <i>BlueBERT</i>	103
4.9.11.	Rangkuman Perbandingan Hasil Percobaan Awal Dan Akhir....	104
4.9.12.	Komparasi Hasil Penelitian dengan Penelitian Terdahulu.....	106
4.9.13.	Komparasi Hasil Dengan Penelitian yang Belum di <i>Fine Tuning</i>	107
BAB V		108
KESIMPULAN DAN SARAN		108
5.1.	Kesimpulan	108
5.2.	Saran.....	109
DAFTAR PUSTAKA		110
LAMPIRAN		115

DAFTAR GAMBAR

Gambar 3. 1 Tahapan Kerangka Kerja	17
Gambar 3. 2 <i>Sample texts</i> pada setiap <i>Dataset</i> NCBI, BC2GM, dan JNLPBA	19
Gambar 3. 3 Visualisasi <i>Box plot</i>	21
Gambar 3. 4 <i>Wordcloud</i> pada setiap <i>dataset</i> NCBI, BC2GM, dan JNLPBA...	22
Gambar 3. 5 <i>Pie chart</i> pada setiap <i>dataset</i> NCBI, BC2GM, dan JNLPBA	23
Gambar 3. 6 Tahapan Data <i>Preprocessing</i>	24
Gambar 3. 7 <i>Text Normalization</i>	25
Gambar 3. 8 <i>Text Tokenization</i>	26
Gambar 3. 9 <i>Word Indexing</i>	26
Gambar 3. 10 <i>Padding</i>	27
Gambar 4. 1 <i>Heatmap</i> Metrik <i>Precision</i> pada <i>Dataset</i> NCBI.....	33
Gambar 4. 2 <i>Heatmap</i> Metrik <i>Recall</i> pada <i>Dataset</i> NCBI.....	35
Gambar 4. 3 <i>Heatmap</i> Metrik <i>F1-Score</i> pada <i>Dataset</i> NCBI	36
Gambar 4. 4 <i>Heatmap</i> Metrik <i>precision</i> pada <i>Dataset</i> BC2GM	38
Gambar 4. 5 <i>Heatmap</i> Metrik <i>recall</i> pada <i>Dataset</i> BC2GM.....	40
Gambar 4. 6 <i>Heatmap</i> Metrik <i>F1-Score</i> pada <i>Dataset</i> BC2GM	41
Gambar 4. 7 <i>Heatmap</i> Metrik <i>Precision</i> pada <i>Dataset</i> JNLPBA.....	44
Gambar 4. 8 <i>Heatmap</i> Metrik <i>Recall</i> pada <i>Dataset</i> JNLPBA.....	46
Gambar 4. 9 <i>Heatmap</i> Metrik <i>F1-Score</i> pada <i>Dataset</i> JNLPBA	49
Gambar 4. 10 Grafik <i>F1-Score</i> dan <i>loss</i> pada masing-masing <i>dataset</i>	51
Gambar 4. 11 <i>Bar Chart Macro average</i> Metrik <i>Precision</i>	55
Gambar 4. 12 <i>Bar Chart Macro average</i> Metrik <i>Recall</i>	57
Gambar 4. 13 <i>Bar Chart Macro average</i> Metrik <i>F1-Score</i>	59
Gambar 4. 14 <i>Bar Chart Macro average</i> Metrik <i>Precision</i>	61
Gambar 4. 15 <i>Bar Chart Macro average</i> Metrik <i>Recall</i>	63
Gambar 4. 16 <i>Bar Chart Macro average</i> metrik <i>F1-Score</i>	65
Gambar 4. 17 <i>Confusion matrix Dataset</i> NCBI	68
Gambar 4. 18 <i>Confusion matrix Dataset</i> BC2GM	70
Gambar 4. 19 <i>Confusion matrix Dataset</i> JNLPBA	73

Gambar 4. 20 <i>Heatmap</i> Metrik <i>Precision</i> pada <i>Dataset</i> NCBI.....	77
Gambar 4. 21 <i>Heatmap</i> Metrik <i>Recall</i> pada <i>Dataset</i> NCBI.....	78
Gambar 4. 22 <i>Heatmap</i> Metrik <i>F1-Score</i> pada <i>Dataset</i> NCBI.....	79
Gambar 4. 23 <i>Heatmap</i> Metrik <i>precision</i> pada <i>Dataset</i> BC2GM.....	81
Gambar 4. 24 <i>Heatmap</i> Metrik <i>Recall</i> pada <i>Dataset</i> BC2GM.....	82
Gambar 4. 25 <i>Heatmap</i> Metrik <i>F1-Score</i> pada <i>Dataset</i> BC2GM	83
Gambar 4. 26 <i>Heatmap</i> Metrik <i>Precision</i> pada <i>Dataset</i> JNLPBA.....	85
Gambar 4. 27 <i>Heatmap</i> Metrik <i>Recall</i> pada <i>Dataset</i> JNLPBA.....	87
Gambar 4. 28 <i>Heatmap</i> Metrik <i>F1-Score</i> pada <i>Dataset</i> JNLPBA	89
Gambar 4. 29 <i>Bar Chart Macro average</i> Metrik <i>Precision</i>	92
Gambar 4. 30 <i>Confusion matrix Dataset</i> NCBI	98
Gambar 4. 31 <i>Confusion matrix Dataset</i> BC2GM.....	100
Gambar 4. 32 <i>Macro average F1-Score</i> Terbaik dari Percobaan Awal dan Percobaan Akhir pada <i>Dataset</i> NCBI, BC2GM, dan JNLPBA	104

DAFTAR TABEL

Tabel 3. 1 Sebaran data.....	18
Tabel 3. 2 keterangan label entitas <i>dataset</i> biomedis	20
Tabel 3. 3 <i>Hyperparameter</i>	29
Tabel 3. 4 <i>Hyperparameter BlueBERT</i>	30
Tabel 4. 1 Skenario percobaan.....	31
Tabel 4. 2 Hasil pengujian Metrik <i>precison</i> pada <i>Dataset</i> NCBI.....	32
Tabel 4. 3 Hasil Pengujian Metrik <i>Recall</i> pada Data <i>Dataset</i> NCBI	34
Tabel 4. 4 Hasil Pengujian Metrik <i>F1-Score</i> pada Data <i>Dataset</i> NCBI	35
Tabel 4. 5 Hasil Pengujian Metrik <i>Precision</i> pada <i>Dataset</i> BC2GM.....	37
Tabel 4. 6 Hasil Pengujian Metrik <i>Recall</i> pada <i>Dataset</i> BC2GM.....	39
Tabel 4. 7 Hasil Pengujian Metrik <i>F1-Score</i> pada <i>Dataset</i> BC2GM.....	40
Tabel 4. 8 Hasil Pengujian Metrik <i>Precision</i> pada <i>Dataset</i> JNLPBA	42
Tabel 4. 9 Hasil pengujian metrik <i>Recall</i> pada <i>Dataset</i> JNLPBA.....	44
Tabel 4. 10 Hasil Pengujian Metrik <i>F1-Score</i> pada <i>Dataset</i> JNLPBA.....	47
Tabel 4. 11 Rangkuman <i>Macro average</i> Metrik <i>Precision</i>	54
Tabel 4. 12 Rangkuman <i>Macro average</i> Metrik <i>Recall</i>	56
Tabel 4. 13 Rangkuman <i>Macro average</i> Metrik <i>F1-Score</i>	58
Tabel 4. 14 Rangkuman <i>Macro average</i> Metrik <i>Precision</i>	60
Tabel 4. 15 Rangkuman <i>Macro average</i> Metrik <i>Recall</i>	62
Tabel 4. 16 Rangkuman <i>Macro average</i> Metrik <i>F1-Score</i>	64
Tabel 4. 17 Durasi Waktu Training Model <i>Biobert</i> dan <i>Clinical BERT</i>	74
Tabel 4. 18 Skenario Percobaan Akhir	75
Tabel 4. 19 Hasil Pengujian Metrik <i>Precision</i> pada <i>Dataset</i> NCBI.....	76
Tabel 4. 20 Hasil Pengujian Metrik <i>Recall</i> pada <i>Dataset</i> NCBI.....	77
Tabel 4. 21 Hasil pengujian Metrik <i>F1-Score</i> pada <i>Dataset</i> NCBI.....	78
Tabel 4. 22 Hasil Pengujian Metrik <i>Precision</i> pada <i>Dataset</i> BC2GM	80
Tabel 4. 23 Hasil pengujian Metrik <i>Recall</i> pada <i>Dataset</i> BC2GM.....	81
Tabel 4. 24 Hasil pengujian Metrik <i>F1-Score</i> pada <i>Dataset</i> BC2GM.....	82
Tabel 4. 25 Hasil Pengujian Metrik <i>Precision</i> pada <i>Dataset</i> JNLPBA	84
Tabel 4. 26 Hasil Pengujian Metrik <i>Recall</i> Pada <i>Dataset</i> JNLPBA.....	86
Tabel 4. 27 Hasil Pengujian Metrik <i>F1-Score</i> pada <i>Dataset</i> JNLPBA.....	88
Tabel 4. 28 Rangkuman <i>Macro average</i> Metrik <i>Precision</i>	90
Tabel 4. 29 Rangkuman <i>Macro average</i> Metrik <i>Recall</i>	92
Tabel 4. 30 Rangkuman <i>Macro average</i> Metrik <i>F1-Score</i>	94
Tabel 4. 31 Durasi Waktu Training Model <i>BioBERT</i> , <i>Clinical BERT</i> , dan <i>BlueBERT</i>	103
Tabel 4. 32 Komparasi Setiap Model pada <i>Dataset</i> Biomedis.....	106
Tabel 4. 33 Komparasi Hasil dengan Penelitian yang Belum di <i>Fine Tuning</i>	107

BAB I

PENDAHULUAN

1.1. Latar Belakang

Dalam era digital yang terus berkembang, digitalisasi data telah menjadi elemen penting dalam berbagai sektor, termasuk bidang kesehatan. Salah satu dampak dari transformasi ini adalah meningkatnya jumlah data biomedis yang tersimpan dalam bentuk digital. Data tersebut mencakup berbagai informasi penting seperti hasil laboratorium, laporan klinis, jurnal ilmiah, hingga catatan observasi pasien, yang seluruhnya memiliki potensi besar untuk mendukung pengambilan keputusan medis dan penelitian lanjutan. Sistem pencatatan elektronik yang sebelumnya dikembangkan dalam konteks rekam medis [1], [2], kini berkembang luas dalam pemanfaatannya untuk menyimpan dan menganalisis berbagai jenis data biomedis. Tantangan utama dari pemanfaatan data biomedis adalah volume yang sangat besar serta formatnya yang cenderung tidak terstruktur, terutama dalam bentuk teks bebas. Hal ini menyebabkan proses pencarian dan ekstraksi informasi penting dari dokumen-dokumen tersebut menjadi sangat kompleks dan memakan waktu bila dilakukan secara manual. Seiring meningkatnya jumlah data, para tenaga medis maupun peneliti menghadapi kesulitan dalam mengekstrak informasi yang relevan secara cepat dan akurat [3]. Oleh karena itu, dibutuhkan pendekatan otomatis yang mampu mengekstrak informasi secara efisien dari teks biomedis.

Salah satu pendekatan yang berkembang pesat adalah *Natural Language Processing* (NLP), cabang dari kecerdasan buatan yang fokus pada interaksi antara komputer dan bahasa manusia [4]. Salah satu komponen penting dalam NLP adalah *Named Entity Recognition* (NER), yaitu teknik yang bertujuan untuk mengenali dan mengklasifikasikan entitas tertentu dari teks, seperti nama penyakit, obat-obatan, prosedur medis, serta gejala [5]. Dalam konteks biomedis, NER memiliki peran krusial dalam mengekstraksi informasi relevan dari literatur ilmiah maupun laporan klinis, sehingga dapat mempercepat proses analisis dan pengambilan keputusan berbasis data. Untuk meningkatkan akurasi dalam tugas NER di bidang biomedis, berbagai model berbasis BERT telah dikembangkan. BERT sendiri merupakan

model bahasa yang mampu memahami konteks kata secara dua arah (*bidirectional*), menjadikannya sangat efektif untuk memahami struktur kalimat yang kompleks [6]. Dari arsitektur dasar ini, lahirlah beberapa varian yang disesuaikan dengan data biomedis, seperti *BioBERT*, yang dilatih menggunakan artikel PubMed dan PMC [7]. *Clinical BERT*, yang dilatih pada data dari catatan kesehatan elektronik seperti MIMIC-III [8]. serta *BlueBERT*, yang menggabungkan data PubMed dan MIMIC-III untuk menangani konteks ilmiah sekaligus klinis.

Transformer memiliki kelebihan dibanding CNN karena mampu menangkap dependensi jarak jauh dalam teks melalui mekanisme *self-attention*, sehingga lebih baik dalam memahami konteks kalimat yang kompleks. Selain itu, *fine-tuning* diperlukan agar model dapat menyesuaikan pengetahuan umum yang sudah dipelajari dengan karakteristik spesifik data tugas target, sehingga meningkatkan akurasi ekstraksi entitas pada domain biomedis.

Penelitian ini berfokus pada analisis dan perbandingan performa ketiga model tersebut dalam tugas *Clinical Named Entity Recognition* (CNER). Model-model ini diuji pada *dataset* relevan yang mencerminkan data biomedis untuk mengevaluasi sejauh mana masing-masing model mampu mengenali dan mengklasifikasikan entitas medis secara akurat. Proses penelitian meliputi pengumpulan data melalui teknik *web scraping*, tahap *preprocessing* data, pelatihan model, dan evaluasi hasil menggunakan metrik evaluasi standar seperti *precision*, *recall*, dan *F1-Score*.

Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam bidang NLP biomedis, khususnya dalam memilih model yang paling optimal untuk implementasi sistem ekstraksi informasi otomatis dari teks medis. Dengan demikian, penelitian ini tidak hanya memberikan gambaran teknis mengenai performa model-model *transformer*, tetapi juga dapat menjadi referensi praktis untuk pengembangan sistem berbasis AI di bidang kesehatan.

1.2. Rumusan Masalah

Berdasarkan pemaparan latar belakang, maka dapat diuraikan perumusan masalah yang terdiri atas :

1. Apakah model BERT yang dilatih secara khusus dengan teks biomedis, seperti *BioBERT*, *Clinical BERT*, dan *BlueBERT* dapat meningkatkan akurasi dan efisiensi dalam tugas NER?
2. Model manakah di antara *BioBERT*, *Clinical BERT*, dan *BlueBERT* yang memberikan hasil terbaik dalam mengenali dan mengklasifikasikan entitas pada teks biomedis berdasarkan metrik evaluasi seperti *precision*, *recall*, dan *F1-Score*?
3. Bagaimana pengaruh variasi parameter pelatihan, seperti *batch size*, terhadap kinerja masing-masing model dalam ekstraksi entitas biomedis?

1.3. Batasan Masalah

Berdasarkan pemaparan latar belakang, maka dapat diuraikan batasan masalah yang terdiri atas:

1. Penelitian ini menggunakan metode *deep learning* berbasis *transformer* untuk menjalankan *Clinical Named Entity Recognition*.
2. Penelitian ini menggunakan model BERT yang dilatih khusus pada teks biomedis, seperti *BioBERT*, *Clinical BERT*, dan *BlueBERT* untuk tugas NER.
3. Penelitian ini akan mengevaluasi kinerja model berdasarkan metrik evaluasi umum dalam NER, seperti *precision*, *recall*, dan *F1-Score*.

1.4. Tujuan Penelitian

Adapun tujuan dari penulisan Tugas Akhir ini, yaitu :

1. Mengembangkan sistem untuk mengidentifikasi dan mengklasifikasikan entitas medis dari teks biomedis tidak terstruktur guna mendukung analisis klinis dan riset kesehatan.
2. Melatih dan mengevaluasi model *BioBERT*, *Clinical BERT*, dan *BlueBERT* dalam memahami terminologi medis untuk tugas *Clinical Named Entity Recognition*.
3. Membandingkan performa ketiga model berdasarkan *precision*, *recall*, dan *F1-Score* untuk menentukan model paling optimal dalam ekstraksi entitas medis.

1.5. Metodologi Penulisan

Penulisan tugas akhir ini menggunakan beberapa pendekatan metodologis, dimulai dari pengumpulan literatur hingga penarikan kesimpulan dan pemberian saran. Setiap metode dijelaskan secara rinci untuk memberikan gambaran menyeluruh mengenai langkah-langkah yang ditempuh dalam penelitian ini.

1.5.1. Metode Studi Pustaka dan Literatur

Pendekatan ini menelusuri dan mengkaji sumber-sumber relevan seperti buku, jurnal, artikel penelitian, dan referensi daring untuk memperoleh pemahaman mendalam tentang teori, metode, dan perkembangan terkini dalam Clinical Named Entity Recognition.

1.5.2. Metode Konsultasi

Metode ini melibatkan diskusi dan konsultasi dengan pihak-pihak yang memiliki keahlian atau pengalaman di bidang yang terkait. Konsultasi dilakukan untuk membantu mengatasi kendala teknis maupun konseptual yang muncul selama proses penyusunan tugas akhir.

1.5.3. Metode Pembuatan Model

Tahapan ini mencakup proses perancangan dan pembuatan model yang disimulasikan sebagai bagian dari eksperimen dalam penelitian. Model yang dibuat diharapkan dapat merepresentasikan permasalahan yang dikaji secara akurat.

1.5.4. Metode Pengujian dan Validasi

Pada tahapan ini, model yang telah dikembangkan akan diuji untuk menilai kinerjanya. Tujuannya adalah memastikan bahwa model tersebut dapat menghasilkan *output* yang sesuai dengan harapan dan dapat diandalkan.

1.5.5. Metode Hasil dan Analisa

Setelah pengujian, hasil akan dianalisis untuk mengevaluasi keunggulan dan keterbatasan model. Analisis ini memberikan gambaran performa serta menjadi dasar pengembangan atau penyesuaian model di masa depan agar lebih optimal.

1.5.6. Metode Penarikan Kesimpulan dan Saran

Bagian akhir dari metodologi ini berfokus pada penyimpulan hasil dari keseluruhan proses penelitian dan menyampaikan rekomendasi yang dapat dijadikan rujukan dalam penelitian lanjutan.

1.6. Sistematika Penulisan

Struktur penulisan tugas akhir ini disusun secara sistematis sebagai berikut:

BAB I PENDAHULUAN

Berisi penjelasan umum mengenai latar belakang masalah, perumusan masalah, batasan ruang lingkup, tujuan dan manfaat penelitian, metodologi yang digunakan, serta sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Menguraikan konsep-konsep dasar, teori-teori terkait, serta hasil-hasil penelitian sebelumnya yang menjadi landasan teoritis penelitian ini.

BAB III METODOLOGI PENELITIAN

Menjelaskan secara terperinci tahapan-tahapan, teknik, serta metode yang diterapkan dalam pelaksanaan penelitian.

BAB IV HASIL DAN ANALISIS

Menyajikan data hasil dari implementasi atau eksperimen yang dilakukan, beserta analisis terhadap temuan tersebut.

BAB V KESIMPULAN

Memuat ringkasan hasil penelitian, kesimpulan yang diperoleh, serta saran untuk penelitian lebih lanjut atau pengembangan ke depan.

DAFTAR PUSTAKA

- [1] J. Lee *et al.*, “*BioBERT: A pre-trained biomedical language representation model for biomedical text mining*,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.
- [2] E. Alsentzer *et al.*, “*Publicly Available Clinical BERT Embeddings*,” 2019. [Online]. Available: <https://www.NCBI.nlm.nih.gov/pmc/>
- [3] W. Handiwidjojo, “REKAM MEDIS ELEKTRONIK.”
- [4] W. V Sujansky and P. Alto, “*The Benefits and Challenges of an Electronic Medical Record: Much More than a ‘Word-Processed’ Patient Chart.*”
- [5] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, “*Natural language processing: An introduction*,” Sep. 2011. doi: 10.1136/amiajnl-2011-000464.
- [6] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, “*Deep learning with word embeddings improves biomedical named entity recognition*,” in *Bioinformatics*, Oxford University Press, Jul. 2017, pp. i37–i48. doi: 10.1093/bioinformatics/btx228.
- [7] W. Zaghouani, I. Vladimir, and M. Ruiz, “*COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter.*” [Online]. Available: <https://github.com/digitalepidemiologylab/covid-twitter-bert>
- [8] A. Y. Chandra, D. Kurniawan, and R. Musa, “*Perancangan Chatbot Menggunakan Dialogflow Natural Language Processing (Studi Kasus: Sistem Pemesanan pada Coffee Shop)*,” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 1, p. 208, Jan. 2020, doi: 10.30865/mib.v4i1.1505.
- [9] X. Zheng, H. Du, X. Luo, F. Tong, W. Song, and D. Zhao, “*BioByGANS: biomedical named entity recognition by fusing contextual and syntactic features through graph attention network in node classification framework*,”

- BMC Bioinformatics*, vol. 23, no. 1, Dec. 2022, doi: 10.1186/s12859-022-05051-9.
- [10] E. Alsentzer *et al.*, “Publicly Available Clinical BERT Embeddings,” Apr. 2019, [Online]. Available: <http://arxiv.org/abs/1904.03323>
 - [11] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo, “Clinical-Longformer and Clinical-BigBird: Transformers for long Clinical sequences,” Jan. 2022, [Online]. Available: <http://arxiv.org/abs/2201.11838>
 - [12] Z. Yuan, Y. Liu, C. Tan, S. Huang, and F. Huang, “Improving Biomedical Pretrained Language Models with Knowledge,” Apr. 2021, [Online]. Available: <http://arxiv.org/abs/2104.10344>
 - [13] F. Nur Rozi and D. Harini Sulistyawati, “KLASIFIKASI BERITA HOAX PILPRES MENGGUNAKAN METODE MODIFIED K-NEAREST NEIGHBOR DAN PEMBOBOTAN MENGGUNAKAN TF-IDF,” 2019.
 - [14] T. Baldwin and Y. Li, “An In-depth Analysis of the Effect of Text Normalization in Social Media.”
 - [15] V. S and J. R, “Text Mining: open Source Tokenization Tools – An Analysis,” *Advanced Computational Intelligence: An International Journal (ACII)*, vol. 3, no. 1, pp. 37–47, Jan. 2016, doi: 10.5121/acii.2016.3104.
 - [16] M. Dwarampudi and N. V. S. Reddy, “Effects of padding on LSTMs and CNNs,” Mar. 2019, [Online]. Available: <http://arxiv.org/abs/1903.07288>
 - [17] “Packing: Towards 2x NLP BERT Acceleration Anonymous.”
 - [18] F. Almeida and G. Xexéo, “Word Embeddings: A Survey.”
 - [19] A. Bakarov, “A Survey of Word Embeddings Evaluation Methods,” 2018.
 - [20] L. G. Birta and G. Arbez, “Modelling and Simulation Exploring Dynamic System Behaviour.”

- [21] L. George and P. Sumathy, “An integrated clustering and BERT framework for improved topic modeling,” *International Journal of Information Technology (Singapore)*, vol. 15, no. 4, pp. 2187–2195, Apr. 2023, doi: 10.1007/s41870-023-01268-w.
- [22] Y. Bengio, I. Goodfellow, and A. Courville, “*Deep Learning*,” 2015.
- [23] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
- [24] L. Wahyunita, “Rekayasa Web Klasifikasi Rocchio pada Data Tidak Terstruktur,” *Jurnal Komunika : Jurnal Komunikasi, Media dan Informatika*, vol. 8, no. 2, p. 88, Dec. 2019, doi: 10.31504/komunika.v8i2.2016.
- [25] B. Maryanto, “BIG DATA DAN PEMANFAATANNYA DALAM BERBAGAI SEKTOR,” 2017.
- [26] M. Kutbi, “Named Entity Recognition Utilized to Enhance Text Classification While Preserving Privacy,” *IEEE Access*, vol. 11, pp. 117576–117581, 2023, doi: 10.1109/ACCESS.2023.3325895.
- [27] K. R. Chowdhary, *Fundamentals of artificial intelligence*. Springer India, 2020. doi: 10.1007/978-81-322-3972-7.
- [28] A. Jain, G. Kulkarni, and V. Shah, “Natural Language Processing,” *International Journal of Computer Sciences and Engineering*, vol. 6, no. 1, pp. 161–167, Jan. 2018, doi: 10.26438/ijcse/v6i1.161167.
- [29] *Proceedings of 2014 2nd International Conference on Technology, Informatics, Management, Engineering, & Environment (TIME-E 2014) : August 19-21, 2014, Bandung, Indonesia*. IEEE, 2014.
- [30] Koroteev MV, “BERT: A Review of Applications in Natural Language Processing and Understanding.”

- [31] J. Lee *et al.*, “BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.
- [32] Y. Zhu, L. Li, H. Lu, A. Zhou, and X. Qin, “Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions,” *J Biomed Inform*, vol. 106, Jun. 2020, doi: 10.1016/j.jbi.2020.103451.
- [33] “Accurate Medical Named Entity Recognition Through Specialized NLP Models.”
- [34] C. Lokker *et al.*, “Deep learning to refine the identification of high-quality Clinical research articles from the biomedical literature: Performance evaluation,” *J Biomed Inform*, vol. 142, Jun. 2023, doi: 10.1016/j.jbi.2023.104384.
- [35] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, “Transformer in Transformer.” [Online]. Available: <https://gitee.com/mindspore/models/>
- [36] M. S. Huang, P. T. Lai, P. Y. Lin, Y. T. You, R. T. H. Tsai, and W. L. Hsu, “Biomedical named entity recognition and linking datasets: Survey and our recent development,” *Brief Bioinform*, vol. 21, no. 6, pp. 2219–2238, Nov. 2020, doi: 10.1093/bib/bbaa054.
- [37] S. Sharma *et al.*, “The NCBI BioCollections Database,” *Database*, vol. 2018, no. 2018, Jan. 2018, doi: 10.1093/database/bay006.
- [38] A. N. Le, H. Morita, and T. Iwakura, “Learning Entity-Likeness with Multiple Approximate Matches for Biomedical NER,” in *International Conference Recent Advances in Natural Language Processing, RANLP*, Incoma Ltd, 2021, pp. 1040–1049. doi: 10.26615/978-954-452-072-4_117.
- [39] S. Raschka, “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning,” Nov. 2018, [Online]. Available: <http://arxiv.org/abs/1811.12808>

- [40] C. Goutte and E. Gaussier, “*A Probabilistic Interpretation of Precision, Recall and A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation,*” 2005. [Online]. Available: <https://www.researchgate.net/publication/221397399>