

***CLINICAL NAMED ENTITY RECOGNITION PADA
DATA BIOMEDIS MENGGUNAKAN MODEL
BIDIRECTIONAL LONG SHORT TERM MEMORY -
CONDITIONAL RANDOM FIELD***

TUGAS AKHIR

**Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer**



OLEH:

TRIA LAILANI

09011282126121

**JURUSAN SISTEM KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA**

2025

LEMBAR PENGESAHAN

SKRIPSI

***CLINICAL NAMED ENTITY RECOGNITION PADA DATA BIOMEDIS
MENGUNAKAN MODEL *BIDIRECTIONAL LONG SHORT TERM
MEMORY-CONDITINIONAL RANDOM FIELD****

Diajukan untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Ilmu Komputer

Oleh :

Tria Lailani

09011282126121

Pembimbing : Dr. Firdaus, S.T., M.Kom.
NIP. 197801212008121003

Mengetahui,

Ketua Jurusan Sistem Komputer



Dr. Ir. Sukemi, M.T.
196612032006041001

AUTHENTICATION PAGE

SKRIPSI

***CLINICAL NAMED ENTITY RECOGNITION ON BIOMEDICAL DATA
USING THE BIDIRECTIONAL LONG SHORT-TERM MEMORY–
CONDITIONAL RANDOM FIELD MODEL***

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

By :

Tria Lailani

09011282126121

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

NIP. 197801212008121003

Approved by,

Head of Computer System Department



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

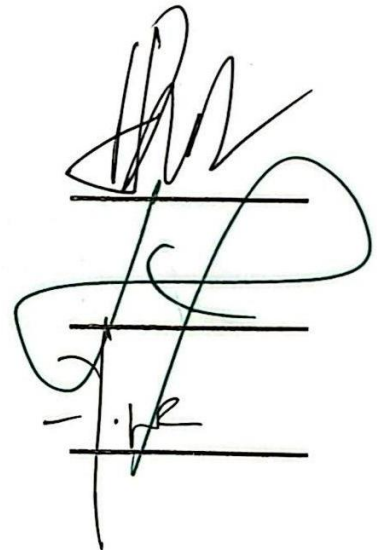
HALAMAN PERSETUJUAN

Telah diuji dan lulus pada :

Hari : Kamis
Tanggal : 26 Juni 2025

Tim Penguji :

1. Ketua Sidang : Prof. Dr. Erwin, S.Si., M.Si.
2. Penguji Sidang : Huda Ubaya, M.T.
3. Pembimbing : Dr. Firdaus, S.T., M.Kom.



Mengetahui, 16/7/25
Ketua Jurusan Sistem Komputer



HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Tria Lailani

Nim : 09011282126102

Judul tugas akhir : *Clinical Named Entity Recognition* Pada Data Biomedis
Menggunakan Model *Bidirectional Long Short Term Memory-
Conditinional Random Field*

Hasil pemeriksaan iThenticate/Turnitin: 6%

Saya menyatakan bahwa laporan tugas akhir ini merupakan hasil kerja saya sendiri dan tidak mengandung unsur penjiplakan atau plagiarisme. Apabila di kemudian hari terbukti terdapat tindakan plagiarisme dalam laporan ini, saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan peraturan yang berlaku.

Pernyataan ini saya buat dengan penuh kesadaran dan tanpa adanya paksaan dari pihak mana pun.



Indralaya, Juli 2025

Yang menyatakan,


METERAI
TEMPEL
47AE7AMX406395769
Tria Lailani

NIM. 09011282126121

HALAMAN PERSEMBAHAN

1. Rasa terima kasih yang tulus saya sampaikan kepada Ayah dan Ibu, yang selalu menjadi sumber kekuatan dan cinta tanpa batas dalam hidup saya. Doa, dukungan, dan semangat yang tak pernah henti telah menemani setiap langkah hingga saya sampai pada titik ini. Skripsi ini saya persembahkan dengan sepenuh hati sebagai bentuk cinta dan penghargaan.
2. Terima kasih telah menjadi bagian berharga dalam setiap langkah hidupku. Kakak, terima kasih atas contoh dan perhatian yang selalu kamu berikan. Kehadiranmu membuatku merasa lebih kuat. Adik, terima kasih atas canda dan semangatmu yang tulus, sering kali menjadi penyemangat di saat aku mulai lelah.

MOTTO

”Berilmulah sebelum kamu berbicara, beramal, atau beraktivitas.”

(HR. Bukhari)

“Just because it burns doesn't mean you're gonna die. You gotta get up and try.”

- P!NK, *“Try”*

“Akan datang hari di mana semua lelah, ragu, dan air mata terbayar tuntas. Bukan karena jalan yang dilalui mudah, melainkan karena pilihan untuk terus bertahan, bukan menyerah. Saat hari itu tiba, rasa syukur akan menyelimuti hati, karena langkah tetap dilanjutkan meski sempat goyah. Setiap usaha yang dilakukan dengan tulus akan berbuah, dan saat itu tiba, tak ada penyesalan”

KATA PENGANTAR

Assalamu'alaikum Warrahmatullahi Wabarakatuh.

Puji dan syukur kepada Allah SWT karena berkat Rahmat dan Karunia-Nya lah sehingga penulis dapat menyelesaikan penyusunan Laporan Tugas Akhir yang berjudul *Clinical Named Entity Recognition Pada Data Biomedis Menggunakan Model Bidirectional Long Short Term Memory-Conditionional Random Field*

Pada kesempatan ini, penulis ingin mengucapkan terima kasih kepada pihak yang telah memberikan bantuan, dorongan, motivasi, semangat, dan bimbingan dalam penyusunan laporan skripsi ini, yakni :

1. Allah Subhanahu Wa Ta'ala, yang telah memberikan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan penulisan skripsi ini dengan baik dan lancar.
2. Keluarga saya yang menjadi salah satu pendukung dalam penulisan tugas akhir, kedua orang tua, kakak perempuan, kakak laki-laki dan adik laki-laki saya. Terima kasih atas segala doa, motivasi, dan dukungan baik secara moril, materil, maupun spiritual.
3. Bapak Prof. Dr. Erwin, S.Si, M.Si., selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
4. Bapak Dr. Ir. Sukemi, M.T. selaku Ketua Jurusan Sistem Komputer Universitas Sriwijaya.
5. Bapak Rahmat Fadli Isnanto, S.Si., M.Sc., selaku Dosen Pembimbing Akademik.
6. Bapak Dr. Firdaus, S.T., M.Kom. selaku Dosen Pembimbing Tugas Akhir.
7. Ibu Prof. Dr. Ir. Siti Nurmaini, M.T., selaku Dosen ISysRG yang turut serta secara membantu dalam mendukung jalannya skripsi ini.
8. Ibu Anggun Islami, M.Kom. selaku mentor divisi data tabular yang telah memberikan bantuan dan masukan dalam penelitian Skripsi ini.

9. Ibu Dr. Ade Iriani Sapitri, M.Kom., Bapak M. Naufal Rahmatullah, Ibu Akhiar Wista Arum, M.Kom. M.T. dan Ibu Dr. Annisa darmawahyuni, M.Kom. selaku mentor yang telah memberikan bantuan dan masukan dalam penelitian skripsi ini.
10. Hazel Aryaputra Anhar, seseorang yang selalu memberi dukungan dan motivasi.
11. Teman seperjuangan, Winda Rahayu, Berliana Ramadhona, Rosa Julia Erizka, Tanti Annastasya, Ririn Purnama Sari, dan Aditiawati, yang selalu memberikan dukungan.
12. Teman-teman ISysRG *batch* 6, khususnya divisi *Text Processing*, M. Azriel Apriadi, Keisyah Syabinatullah, Indah Gala Putri, dan Tiara Oktarina.
13. Teman-teman Sistem Komputer Angkatan 2021

Penulis menyadari sepenuhnya bahwa laporan tugas akhir ini masih jauh dari kata sempurna, oleh karena itu penulis mengharapkan saran dan kritik dari semua pihak yang berkepentingan demi laporan yang lebih baik lagi. Penulis berharap semoga laporan tugas akhir ini dapat bermanfaat serta dapat memberikan pengetahuan dan wawasan bagi semua pihak yang membutuhkannya.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Indralaya, Juli 2025



Tria Lailani

NIM 09011282126121

CLINICAL NAMED ENTITY RECOGNITION PADA DATA BIOMEDIS MENGUNAKAN MODEL *BIDIRECTIONAL LONG SHORT TERM MEMORY CONDITIONAL RANDOM FIELD*

Tria Lailani (09011282126121)

Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya

Email: lailanitria@gmail.com

ABSTRAK

Kemajuan dalam bidang *Natural Language Processing* (NLP) telah mendorong peningkatan signifikan dalam ekstraksi informasi dari teks biomedis yang tidak terstruktur, khususnya dalam mengenali entitas bernama klinis seperti penyakit, gen, dan protein. Penelitian ini bertujuan untuk mengevaluasi kinerja model *Bidirectional Long Short-Term Memory* (Bi-LSTM) dan Bi-LSTM dengan *Conditional Random Fields* (Bi-LSTM-CRF) untuk tugas *Clinical Named Entity Recognition* (CNER). Tiga dataset standar digunakan dalam penelitian ini, yaitu NCBI-Disease, BC2GM, dan JNLPBA. Selain itu, penelitian ini juga menyelidiki pengaruh integrasi *word embedding* GloVe terhadap kinerja model. Hasil evaluasi menunjukkan bahwa model Bi-LSTM secara umum memiliki performa yang lebih baik dalam hal *precision* dan *recall*, sementara Bi-LSTM-CRF menunjukkan keunggulan dalam menjaga konsistensi pelabelan. Evaluasi dilakukan dengan menggunakan metrik *precision*, *recall*, dan *F1-Score*. Temuan dari penelitian ini diharapkan dapat mendukung pengembangan model CNER yang lebih efisien untuk sistem informasi medis serta meningkatkan kualitas pengambilan keputusan klinis.

Kata Kunci : Pengenalan Entitas Klinis, Bi-LSTM, CRF, GloVe, Teks Biomedis, NLP

CLINICAL NAMED ENTITY RECOGNITION ON BIOMEDICAL DATA USING BIDIRECTIONAL LONG SHORT TERM MEMORY CONDITIONAL RANDOM FIELD MODEL

Tria Lailani (09011282126121)

*Department of Computer System, Faculty of Computer Science,
Sriwijaya University
Email: lailanitria@gmail.com*

ABSTRACT

Advancements in Natural Language Processing (NLP) have improved the extraction of information from unstructured biomedical text, particularly in recognizing clinical named entities like diseases, genes, and proteins. This study evaluates the performance of Bi-LSTM and Bi-LSTM-CRF models for Clinical Named Entity Recognition (CNER) using three benchmark datasets: NCBI-Disease, BC2GM, and JNLPBA. It also investigates the effect of integrating GloVe word embeddings. Results show that Bi-LSTM generally outperforms Bi-LSTM-CRF in precision and recall, while Bi-LSTM-CRF maintains better label consistency. Evaluation is based on precision, recall, and F1-score, with findings supporting the development of more effective CNER models for clinical applications.

Keywords : *Clinical Named Entity Recognition, Bi-LSTM, CRF, GloVe, Biomedical Text, NLP.*

DAFTAR ISI

LEMBAR PENGESAHAN	ii
AUTHENTICATION PAGE	iii
HALAMAN PERSETUJUAN.....	iv
HALAMAN PERNYATAAN.....	v
HALAMAN PERSEMBAHAN.....	vi
KATA PENGANTAR.....	vii
ABSTRAK	ix
ABSTRACT	x
DAFTAR ISI.....	xi
DAFTAR GAMBAR.....	xvi
DAFTAR TABEL.....	xviii
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	2
1.3. Batasan Masalah.....	2
1.4. Tujuan.....	3
1.5. Metodologi Penulisan.....	3
1.5.1. Metode Studi Pustaka dan Literatur	3
1.5.2. Metode Konsultasi.....	3
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.....	4

1.6. Sistematika Penulisan.....	4
BAB II TINJAUAN PUSTAKA.....	6
2.1. Penelitian Terdahulu	6
2.2. Data Biomedis.....	7
2.2.1. NCBI- <i>Disease</i>	7
2.2.2. BC2GM	7
2.2.3. JNLPBA	8
2.3. <i>Text Pre-processing</i>	8
2.3.1. <i>Text Normalization</i>	8
2.3.2. <i>Text Tokenization</i>	9
2.3.3. <i>Word Indexing</i>	9
2.4. <i>Natural Language Processing (NLP)</i>	9
2.5. <i>Named Entity Recognition (NER)</i>	10
2.6. <i>Clinical Named Entity Recognition (CNER)</i>	10
2.7. <i>Deep Learning</i>	10
2.8. Bi-LSTM.....	11
2.9. CRF	12
2.10. <i>Word Embedding</i>	12
2.10.1. GloVe	13
2.11. Pengukuran Kinerja.....	13
2.11.1. <i>Confusion Matrix</i>	13
2.11.2. <i>Recall</i>	14
2.11.3. <i>Precision</i>	14

2.11.4. F1-Score	14
BAB III METODOLOGI PENELITIAN	15
3.1. Kerangka Kerja	15
3.2. Akuisisi Data.....	16
3.3. <i>Data Splitting</i>	17
3.4. <i>Exploratory Data Analysis</i> (EDA).....	18
3.4.1. <i>Descriptive Analysis</i>	18
3.4.2. <i>Visualization of Entity Label Distribution</i>	18
3.4.3. <i>Visualization of Entity Label Percentage</i>	19
3.4.5. <i>Wordcloud Visualization Of Dominan Words</i>	20
3.5. <i>Data Pre-Processing</i>	21
3.5.1. <i>Text Normalization</i>	21
3.5.2. <i>Text Tokenization</i>	22
3.5.3. <i>Word Indexing</i>	23
3.5.4. <i>Padding</i>	24
3.6. Pelatihan Model	25
3.6.1. Bi-LSTM	25
3.6.2. Bi-LSTM-CRF.....	26
3.7. Penambahan <i>Word Embedding</i> sebagai Komparasi.....	28
3.8. Evaluasi Model.....	31
BAB IV HASIL DAN PEMBAHASAN.....	32
4.1. Skenario Percobaan	32
4.2. Hasil Percobaan pada Dataset NCBI- <i>Disease</i>	33

4.3. Hasil Percobaan pada Dataset BC2GM.....	39
4.4. Hasil Percobaan pada Dataset JNLPBA.....	45
4.5. <i>Learning Curve</i> pada Data Latih dan Validasi.....	54
4.6. <i>Macro Average</i> Dataset NCBI- <i>Disease</i> , BC2GM dan JNLPBA	57
4.6.1. <i>Macro Average</i> Ketiga Dataset Biomedis dengan <i>Batch Size</i> 16.....	57
4.6.2. <i>Macro Average</i> Ketiga Dataset Biomedis dengan <i>Batch Size</i> 32.....	65
4.7. <i>Confusion Matrix</i> Pada Data Uji	72
4.7.1. <i>Confusion Matrix</i> Pada Dataset NCBI- <i>Disease</i>	72
4.7.2. <i>Confusion matrix</i> pada Dataset BC2GM.....	75
4.7.3. <i>Confusion Matrix</i> pada Dataset JNLPBA.....	78
4.8. Durasi Waktu Pelatihan Model Bi-LSTM dan Bi-LSTM-CRF	81
4.9. Peningkatan Metrik Evaluasi Model Pada Ketiga Dataset Biomedis	82
4.9.1. Skenario Percobaan Akhir	82
4.9.2. Hasil Percobaan pada Dataset NCBI- <i>Disease</i>	83
4.9.3. Hasil Percobaan pada Dataset BC2GM.....	90
4.9.3. Hasil Percobaan pada Dataset JNLPBA.....	96
4.10. <i>Learning Curve</i> Data Latih dan Validasi	105
4.11. <i>Macro Average</i> pada Ketiga Data Biomedis	109
4.12. <i>Confusion Matrix</i> pada Data Uji	118
4.12.1. <i>Confusion Matrix</i> pada Data Uji NCBI- <i>Disease</i>	118
4.12.2. <i>Confusion Matrix</i> pada Data Uji BC2GM.....	121
4.12.3. <i>Confusion Matrix</i> pada Data Uji JNLPBA.....	123
4.13. Rangkuman Waktu Latih pada Ketiga Model.....	126

4.14. Rangkuman Perbandingan Hasil Percobaan Awal dan Akhir.....	127
4.15. Komparasi Hasil Penelitian.....	128
BAB V KESIMPULAN DAN SARAN	130
5.1. Kesimpulan	130
5.2. Saran.....	131
DAFTAR PUSTAKA	132
LAMPIRAN.....	136

DAFTAR GAMBAR

Gambar 2.1 Ilustrasi <i>Text Tokenization</i> [15].....	9
Gambar 2.2 Ilustrasi Arsitektur Bi-LSTM [23].	11
Gambar 3.1 Rancangan Alur Kerja.....	16
Gambar 3.2 Perbandingan Panjang Kata Setiap Dataset	19
Gambar 3.3 Distribusi Label Entitas.....	20
Gambar 3.4 <i>Word Cloud</i>	21
Gambar 3.5 <i>Text Normalization</i>	22
Gambar 3.6 <i>Text Tokenization</i>	23
Gambar 3.7 <i>Word Indexing</i>	24
Gambar 3.8 <i>Padding</i>	25
Gambar 3.9 Arsitektur Bi-LSTM.....	26
Gambar 3.10 Arsitektur Bi-LSTM-CRF	27
Gambar 3.11 Arsitektur GloVe Bi-LSTM	29
Gambar 3.12 Arsitektur GloVe Bi-LSTM-CRF	30
Gambar 4.1 <i>Heatmap Precision</i> NCBI-Disease	34
Gambar 4.2 <i>Heatmap Recall</i> NCBI-Disease	36
Gambar 4.3 <i>Heatmap F1-Score</i> NCBI-Disease.....	38
Gambar 4.4 <i>Heatmap Precision</i> BC2GM.....	40
Gambar 4.5 <i>Heatmap Recall</i> BC2GM	42
Gambar 4.6 <i>Heatmap F1-Score</i> BC2GM.....	44
Gambar 4.7 <i>Heatmap Precision</i> JNLPBA	47
Gambar 4.9 <i>Heatmap F1-Score</i> JNLPBA.....	53
Gambar 4.10 <i>Learning Curve</i> Terbaik Dari Ketiga Dataset Pada Data Latih Dan Validasi	56
Gambar 4.11 <i>Bar Chart Macro Average</i> Metrik <i>Precision</i> Pada A) Dataset NCBI-Disease, B) Dataset BC2GM, C) Dataset JNLPBA.....	60

Gambar 4.12 <i>Bar Chart Macro Average Metrik Recall</i> Pada a) Dataset NCBI-Disease, b) Dataset BC2GM, c) Dataset JNLPBA	62
Gambar 4.13 <i>Bar Chart Macro Average Metrik F1-Score</i> Pada A) Dataset NCBI-Disease, B) Dataset BC2GM, C) Dataset JNLPBA.....	64
Gambar 4.14 <i>Bar Chart Macro Average Precision</i> Pada A) Dataset NCBI-Disease, B) Dataset BC2GM, C) Dataset JNLPBA	67
Gambar 4.15 <i>Bar Chart Macro Average Recall</i> Pada A) Dataset NCBI-Disease, B) Dataset BC2GM, C) Dataset JNLPBA	69
Gambar 4.16 <i>Bar Chart Macro Average F1-Score</i> Pada A) Dataset NCBI-Disease, B) Dataset BC2GM, C) Dataset JNLPBA	71
Gambar 4.17 <i>Confusion matrix</i> Dataset NCBI-Disease	74
Gambar 4.18 <i>Confusion Matrix</i> Dataset BC2GM.....	77
Gambar 4.19 <i>Confusion matrix</i> Dataset JNLPBA	80
Gambar 4.20 <i>Heatmap Precision</i> NCBI-Disease	85
Gambar 4.21 <i>Heatmap Recall</i> NCBI-Disease	87
Gambar 4.22 <i>Heatmap F1-Score</i> NCBI-Disease	89
Gambar 4.23 <i>Heatmap Precision</i> BC2GM.....	91
Gambar 4.24 <i>Heatmap Recall</i> BC2GM	93
Gambar 4.26 <i>Heatmap Precision</i> JNLPBA	98
Gambar 4.27 <i>Heatmap Recall</i> JNLPBA	101
Gambar 4.28 <i>Heatmap F1-Score</i> JNLPBA.....	104
Gambar 4.29 <i>Learning Curve</i> Dari Ketiga Dataset.....	109
Gambar 4.30 <i>Bar Chart Macro Average Metrik Precision</i>	112
Gambar 4.31 <i>Bar Chart Macro Average Metrik Recall</i>	118
Gambar 4.32 <i>Confusion Matrix</i> Pada Data Uji NCBI-Disease.....	120
Gambar 4.33 <i>Confusion Matrix</i> Pada Data Uji BC2GM	123
Gambar 4.34 <i>Confusion Matrix</i> pada Data JNLPBA.....	125
Gambar 4.35 <i>Bar Chart</i> Perbandingan Hasil Percobaan Awal dan Akhir.....	127

DAFTAR TABEL

Tabel 2.1 <i>Confusion Matrix</i>	14
Tabel 3.1 Persentase Dan Pembagian Data.....	17
Tabel 3.2 Keterangan Label Entitas	18
Tabel 3.3 <i>Hyperparameter</i> Pada Penelitian	28
Tabel 3.4 <i>Hyperparameter</i> Pada Model Komparasi	31
Tabel 4.1 Skenario Percobaan.....	32
Tabel 4.2 Hasil Pengujian Pada Data NCBI- <i>Disease</i> Metrik <i>Precision</i>	33
Tabel 4.3 Hasil Pengujian Pada Data NCBI- <i>Disease</i> Metrik <i>Recall</i>	35
Tabel 4.4 Hasil Pengujian Pada Data NCBI- <i>Disease</i> Metrik <i>F1-Score</i>	37
Tabel 4.5 Hasil Percobaan Dataset BC2GM Dengan Metrik <i>Precision</i>	39
Tabel 4.6 Hasil Percobaan Dataset BC2GM Dengan Metrik <i>Recall</i>	41
Tabel 4.7 Hasil Percobaan Dataset BC2GM Dengan Metrik <i>F1-Score</i>	43
Tabel 4.8 Hasil Percobaan Dataset JNLPBA Dengan Metrik <i>Precision</i>	45
Tabel 4.9 Hasil Percobaan Dataset JNLPBA Dengan Metrik <i>Recall</i>	48
Tabel 4.10 Hasil Percobaan Dataset JNLPBA Dengan Metrik <i>F1-Score</i>	51
Tabel 4.11 <i>Macro Average</i> Metrik <i>Precision</i>	58
Tabel 4.13 <i>Macro Average</i> Metrik <i>F1-Score</i>	62
Tabel 4.14 <i>Macro Average</i> Metrik <i>Precision</i>	65
Tabel 4.17 Durasi Waktu Pelatihan	81
Tabel 4.18 Skenario Percobaan Akhir	83
Tabel 4.19 Hasil Percobaan Pada Dataset NCBI- <i>Disease</i> Dengan Metrik <i>Recall</i>	84
Tabel 4.20 Hasil Percobaan Pada Dataset NCBI- <i>Disease</i> Dengan Metrik <i>Recall</i>	86
Tabel 4.21 Hasil Percobaan Pada Dataset NCBI- <i>Disease</i> Metrik <i>F1-Score</i>	88
Tabel 4.22 Hasil Percobaan Pada Dataset BC2GM Dengan Metrik <i>Precision</i>	90
Tabel 4.23 Hasil Percobaan Pada Dataset BC2GM Dengan Metrik <i>Recall</i>	92
Tabel 4.24 Hasil Percobaan Pada Dataset BC2GM Dengan Metrik <i>F1-Score</i>	94

Tabel 4.24 Hasil Percobaan Pada Dataset JNLPBA Dengan Metrik <i>Precision</i>	96
Tabel 4.25 Hasil Percobaan Pada Dataset JNLPBA Dengan Metrik <i>Recall</i>	99
Tabel 4.26 Hasil Percobaan Pada Dataset JNLPBA Dengan Metrik <i>F1-Score</i>	102
Tabel 4.27 <i>Macro Average</i> Metrik <i>Precision</i>	110
Tabel 4.28 <i>Macro Average</i> Metrik <i>Recall</i>	112
Tabel 4.29 <i>Macro Average</i> Metrik <i>F1-Score</i>	115
Tabel 4.30 Rangkuman Durasi Pelatihan.....	126
Tabel 4.31 Komparasi Hasil Penelitian.....	129

BAB I

PENDAHULUAN

1.1. Latar Belakang

Perkembangan pesat teknologi informasi di bidang kesehatan telah mendorong lonjakan besar dalam ketersediaan data biomedis. Catatan medis, hasil pemeriksaan laboratorium, laporan radiologi, dan data genomik adalah beberapa contoh data biomedis yang tersimpan dalam format digital [1]. Data ini sangat berguna dalam mendukung pengambilan keputusan klinis dan pemantauan penyakit, membantu tenaga medis untuk membuat keputusan yang lebih baik dan lebih cepat dalam perawatan pasien [2]. Namun, salah satu tantangan utama yang dihadapi dalam pengolahan data biomedis adalah banyak teks klinis yang tidak terstruktur. Teks yang tidak terstruktur ini membuat penemuan pengetahuan dan aplikasi skala besar dari data medis menjadi sulit dan terbatas [3].

Untuk mengatasi masalah ini, salah satu teknik *Natural Language Processing* (NLP), seperti *Clinical Named Entity Recognition* (CNER) digunakan untuk mengubah data teks medis asli menjadi informasi terstruktur. Dalam data biomedis, teknik CNER mencari dan mengklasifikasikan istilah klinis seperti penyakit, gejala, obat, prosedur, atau bagian tubuh dalam teks, sehingga informasi klinis yang awalnya tersembunyi dalam teks bebas dapat diekstraksi dan disusun ulang ke dalam bentuk yang lebih terstruktur [1]. Hasil entitas berlabel ini yang dihasilkan oleh CNER sangat penting untuk membangun basis data klinis yang unggul dan sistem pendukung keputusan medis [4].

Penelitian ini bertujuan untuk mengembangkan dan mengimplementasikan metode NLP dan *Machine Learning* dengan menggunakan model *Bidirectional Long Short Term Memory-Conditional Random Field* (Bi-LSTM-CRF). Model ini dirancang untuk secara efektif mengekstraksi informasi berguna dari data biomedis, sehingga mendukung pengolahan data yang lebih terstruktur dan memberikan kontribusi terhadap peningkatan kualitas perawatan kesehatan secara keseluruhan [5]. Penelitian ini diharapkan dapat memberikan kontribusi terhadap peningkatan

kualitas layanan kesehatan dengan memanfaatkan teknologi terkini dalam bidang NLP.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, penelitian ini akan merumuskan beberapa permasalahan yang perlu dijawab, antara lain :

1. Bagaimana membangun dan melatih model CNER berbasis arsitektur Bi-LSTM dan Bi-LSTM-CRF untuk mengekstraksi entitas klinis dari data biomedis tidak terstruktur?
2. Bagaimana pengaruh integrasi *word embedding* GloVe terhadap kemampuan model Bi-LSTM dan Bi-LSTM-CRF dalam mengenali entitas seperti penyakit, gen, dan protein?
3. Bagaimana perbandingan performa model Bi-LSTM dan Bi-LSTM-CRF dalam mengekstraksi entitas klinis berdasarkan metrik evaluasi seperti *precision*, *recall*, dan *F1-Score*?
4. Bagaimana kinerja model pada masing-masing dataset biomedis?

1.3. Batasan Masalah

Dalam penelitian ini, beberapa batasan masalah ditetapkan untuk memperjelas ruang lingkup penelitian, yaitu:

1. Penelitian ini hanya difokuskan pada tugas CNER untuk mengekstraksi entitas medis seperti penyakit, gen, dan protein dari data biomedis tidak terstruktur.
2. Model yang digunakan terbatas pada Bi-LSTM dan Bi-LSTM-CRF, tanpa membandingkan dengan model konvensional lain.
3. Penelitian ini menggunakan representasi kata GloVe sebagai *embedding* untuk mendukung pemahaman semantik kata oleh model.
4. Dataset yang digunakan hanya terbatas pada data NCBI-*Disease*, BC2GM dan JNLPBA, yang merupakan dataset biomedis beranotasi yang tersedia secara publik.
5. Evaluasi kinerja model dibatasi pada penggunaan metrik *precision*, *recall*, dan *F1-Score* untuk menilai efektivitas dalam mengenali entitas klinis.

1.4. Tujuan

Tujuan yang akan dicapai dalam penelitian ini adalah :

1. Mengembangkan dan menerapkan model Bi-LSTM dan Bi-LSTM-CRF untuk tugas CNER pada data biomedis.
2. Mengintegrasikan *word embedding* ke dalam model Bi-LSTM dan Bi-LSTM-CRF untuk meningkatkan representasi semantik kata dalam proses ekstraksi entitas klinis.
3. Menganalisis dan membandingkan performa Bi-LSTM dan Bi-LSTM-CRF dalam mengekstraksi entitas klinis dari teks tidak terstruktur berdasarkan *precision*, *recall*, dan *F1-Score*.
4. Mengevaluasi kinerja model pada berbagai dataset biomedis seperti NCBI-Disease, BC2GM, dan JNLPBA guna mengetahui tingkat keefektifan model pada karakteristik data yang berbeda.

1.5. Metodologi Penulisan

Metode yang digunakan dalam penulisan tugas akhir ini, mulai dari studi pustaka hingga penarikan kesimpulan dan saran. Setiap metode akan diuraikan secara rinci untuk memberikan gambaran lengkap mengenai pendekatan yang diambil dalam penelitian ini.

1.5.1. Metode Studi Pustaka dan Literatur

Metode ini dilakukan dengan cara mencari dan mengumpulkan referensi data berupa *literature* yang terdapat pada buku, paper dan internet.

1.5.2. Metode Konsultasi

Metode ini dilakukan dengan cara berkonsultasi kepada ahli atau seseorang yang memiliki pengetahuan dan pengalaman dalam mengatasi masalah yang dihadapi selama penulisan tugas akhir.

1.5.3. Metode Pembuatan Model

Metode ini melibatkan perancangan dan pengembangan suatu model melalui penggunaan simulasi.

1.5.4. Metode Pengujian dan Validasi

Metode ini menjelaskan bagaimana menguji model yang telah dibuat, dan menentukan apakah suatu model yang dibuat tersebut memberikan hasil yang baik.

1.5.5. Metode Hasil dan Analisa

Metode ini merupakan hasil dalam penelitian akan dianalisis untuk mengevaluasi kelebihan dan kekurangan model.

1.5.6. Metode Penarikan Kesimpulan dan Saran

Metode ini memberikan kesimpulan hasil penelitian yang telah dilakukan dan memberikan rekomendasi yang dapat dijadikan referensi untuk penelitian berikutnya.

1.6. Sistematika Penulisan

Sistematika penulisan yang digunakan dalam tugas akhir ini memiliki beberapa struktur penulisan, yaitu :

BAB I PENDAHULUAN

Pada bab ini, penulis menjelaskan mengenai maksud dan tujuan penelitian ini yang mencakup latar belakang, tujuan penelitian, manfaat penelitian, batasan masalah, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Pada bab ini, penulis yang lakukan adalah mengumpulkan informasi yang relevan dengan kasus yang diangkat dari berbagai macam sumber sebagai bahan referensi.

BAB III METODELOGI PENELITIAN

Pada bab ini, penulis akan menjelaskan mengenai proses penelitian yang dilakukan, mulai dari persiapan data hingga sampai pada tahap analisis dan kesimpulan.

BAB IV HASIL DAN ANALISIS

Pada bab ini, penulis akan menyajikan penjelasan mengenai analisis dan hasil evaluasi yang didapat setelah melakukan penelitian.

BAB V KESIMPULAN DAN SARAN

Pada bab ini, penulis akan memberikan kesimpulan dari hasil penelitian yang dilakukan serta memberikan saran dari hasil yang diperoleh agar sebagai bahan referensi orang lain yang melakukan penelitian dibidang yang sama.

DAFTAR PUSTAKA

Daftar pustaka berisi daftar refrensi dari sumber-sumber informasi yang digunakan dalam metode literatur.

LAMPIRAN

Lampiran mencakup formulir perbaikan dan juga pemeriksaan tingkat kemiripan karya dengan sumber lain.

DAFTAR PUSTAKA

- [1] M. C. Durango, E. A. Torres-Silva, and A. Orozco-Duque, “Named Entity Recognition in Electronic Health Records: A Methodological Review,” *Healthc. Inform. Res.*, vol. 29, no. 4, pp. 286–300, 2023, doi: 10.4258/hir.2023.29.4.286.
- [2] Y. W. Nugroho and F. A. Pramudita, “Peran Rekam Medis Elektronik dalam Meningkatkan Efisiensi, Kualitas Layanan Kesehatan, dan Keselamatan Perawatan Pasien: Analisis Systematic Literature Review,” *Media Publ. Promosi Kesehat. Indones.*, vol. 7, no. 2, pp. 343–350, 2024, doi: 10.56338/mppki.v7i2.4867.
- [3] S. Darmiani, B. Y. Pratama, B. Islamy, and T. A. Hidayat, “Tantangan Integrasi Rekam Medis Elektronik dengan Sistem Manajemen Rumah Sakit : Dampak pada Keamanan Data dan Efisiensi Biaya Operasional-A Systematic Review Universitas Sangga Buana YPKP , Indonesia Tantangan Integrasi Rekam Medis Elektronik dengan Sist,” vol. 4, no. November, pp. 1107–1116, 2024.
- [4] I. Keraghel, S. Morbieu, and M. Nadif, “A survey on recent advances in named entity recognition,” pp. 1–42, 2024, [Online]. Available: <http://arxiv.org/abs/2401.10825>
- [5] S. H. Permatasari, I. M. Nur, and F. Fauzi, “Metode Bidirectional Long Short-Term Memory (BiLSTM) Untuk Memprediksi Harga Saham BBRI Dengan Optimasi Nesterov Adaptive Moment (Nadam),” pp. 1151–1159.
- [6] 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, pp. 1–19, 2022, doi: 10.1186/s12859-022-05051-9.

- [7] G. Ning and Y. Bai, “Biomedical named entity recognition based on Glove-BLSTM-CRF model,” *J. Comput. Methods Sci. Eng.*, vol. 21, no. 1, pp. 125–133, 2021, doi: 10.3233/JCM-204419.
- [8] L. Luo *et al.*, “An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition,” *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, 2018, doi: 10.1093/bioinformatics/btx761.
- [9] N. Goyal and N. Singh, “Named entity recognition and relationship extraction for biomedical text: A comprehensive survey, recent advancements, and future research directions,” *Neurocomputing*, vol. 618, no. December 2024, p. 129171, 2025, doi: 10.1016/j.neucom.2024.129171.
- [10] H. Ur Rahman, T. Hahn, and R. Segall, “Biomedical Disease name entity recognition using NCBI-DISEASE corpus,” *CEUR Workshop Proc.*, vol. 1747, 2016.
- [11] L. Smith *et al.*, “Overview of BioCreative II gene mention recognition,” *Genome Biol.*, vol. 9, no. SUPPL. 2, pp. 1–19, 2008, doi: 10.1186/gb-2008-9-s2-s2.
- [12] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, “Introduction to the bio-entity recognition task at JNLPBA,” p. 70, 2004, doi: 10.3115/1567594.1567610.
- [13] D. Rifaldi, Abdul Fadlil, and Herman, “Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet ‘Mental Health,’” *Decod. J. Pendidik. Teknol. Inf.*, vol. 3, no. 2, pp. 161–171, 2023, doi: 10.51454/decode.v3i2.131.
- [14] A. J. S. Berkowitz, A. Srinivasan, J. Miguel, and A. Cortina, “Biomedical Text Normalization through Generative Modeling,” 2025.
- [15] A. Pimpalkar and J. R. Raj R, “MBiLSTMGloVe: Embedding GloVe knowledge into the corpus using multi-layer BiLSTM deep learning model for social media sentiment analysis,” *Expert Syst. Appl.*, vol. 203, no. December 2021, p. 117581, 2022, doi: 10.1016/j.eswa.2022.117581.

- [16] P. López-Úbeda, M. C. Díaz-Galiano, L. A. Ureña-López, and M. T. Martín-Valdivia, “Combining word embeddings to extract chemical and drug entities in biomedical literature,” *BMC Bioinformatics*, vol. 22, no. 1, pp. 1–18, 2021, doi: 10.1186/s12859-021-04188-3.
- [17] D. Jurafsky and J. H. Martin, “Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models Third Edition draft Summary of Contents,” 2024.
- [18] B. Zhou, G. Yang, Z. Shi, and S. Ma, “Natural Language Processing for Smart Healthcare,” *IEEE Rev. Biomed. Eng.*, vol. 17, pp. 4–18, 2024, doi: 10.1109/RBME.2022.3210270.
- [19] Y. Wen, C. Fan, G. Chen, X. Chen, and M. Chen, “A Survey on Named Entity Recognition,” *Lect. Notes Electr. Eng.*, vol. 571 LNEE, no. 1991, pp. 1803–1810, 2020, doi: 10.1007/978-981-13-9409-6_218.
- [20] H. Tu, L. Han, and G. Nenadic, “Extraction of Medication and Temporal Relation from Clinical Text using Neural Language Models,” *Proc. - 2023 IEEE Int. Conf. Big Data, BigData 2023*, pp. 2735–2744, 2023, doi: 10.1109/BigData59044.2023.10386489.
- [21] N. Nath, S. H. Lee, and I. Lee, “NEAR: Named entity and attribute recognition of clinical concepts,” *J. Biomed. Inform.*, vol. 130, 2022, doi: 10.1016/j.jbi.2022.104092.
- [22] Q. Wei *et al.*, “A study of deep learning approaches for medication and adverse drug event extraction from clinical text,” *J. Am. Med. Informatics Assoc.*, vol. 27, no. 1, pp. 13–21, 2020, doi: 10.1093/jamia/ocz063.
- [23] E. da S. Farias and E. P. Junior, “Application of Attention Mechanism with Bidirectional Long Short-Term Memory (BiLSTM) and CNN for Human Conflict Detection using Computer Vision,” 2025, [Online]. Available: <http://arxiv.org/abs/2502.18555>

- [24] F. Shahrokh, N. Ghadiri, R. Samani, and M. Moradi, “Multi-level biomedical NER through multi-granularity embeddings and enhanced labeling,” pp. 1–22, 2023.
- [25] P. M. Brennan, J. J. M. Loan, N. Watson, P. M. Bhatt, and P. A. Bodkin, “Pre-operative obesity does not predict poorer symptom control and quality of life after lumbar disc surgery,” *Br. J. Neurosurg.*, vol. 31, no. 6, pp. 682–687, 2017, doi: 10.1080/02688697.2017.1354122.
- [26] A. Suragala, P. Venkateswarlu, and M. China Raju, “A Comparative Study of Performance Metrics of Data Mining Algorithms on Medical Data,” *Lect. Notes Electr. Eng.*, vol. 698, no. February, pp. 1549–1556, 2021, doi: 10.1007/978-981-15-7961-5_139.
- [27] J. Zhang, L. Xiao, Y. Zhang, J. Lai, and Y. Yang, “Optimization and Performance Evaluation of Deep Learning Algorithm in Medical Image Processing,” *Front. Comput. Intell. Syst.*, vol. 7, no. 3, pp. 67–71, 2024, doi: 10.54097/de0qx980.
- [28] A. Rammal, E. Yammine, W. Fahs, and R. Khatoun, “Evaluation of Supervised machine learning classifiers for detecting the degrees of possibility of Coronavirus Disease infection,” *ResearchSquare*, 2021, [Online]. Available: <http://www.epistemonikos.org/documents/81901dc3692c327453e0399af50895dfd6a4122b>
- [29] S. Singh and A. Kumar, “Review of Methods for Handling Class-Imbalanced in Classification Problems”.