

# ***CLINICAL NAMED ENTITY RECOGNITION PADA DATA BIOMEDIS MENGGUNAKAN MODEL BERT***

**SKRIPSI**

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



**OLEH:**

**KEISYAH SABINATULLAH QUR'AINI**

**09011182126011**

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

**2025**

# HALAMAN PENGESAHAN

## SKRIPSI

### ***CLINICAL NAMED ENTITY RECOGNITION* PADA DATA BIOMEDIS MENGGUNAKAN MODEL BERT**

Diajukan Untuk Melengkapi Salah Satu Syarat  
Memperoleh Gelar Sarjana Komputer

Oleh:

**Keisyah Sabinatullah Qur'aini**

**09011182126011**

**Pembimbing 1 : Dr. Firdaus, S.T., M.Kom.**  
**NIP. 197801212008121003**

**Pembimbing 2 : Prof. Dr. Ir. Bambang Tutuko, M.T.**  
**NIP. 196001121989031002**

**Mengetahui**

**Ketua Jurusan Sistem Komputer**



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

**AUTHENTICATION PAGE**

**SKRIPSI**

***CLINICAL NAMED ENTITY RECOGNITION ON  
BIOMEDICAL DATA USING BERT MODEL***

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

*By:*

**Keisyah Sabinatullah Qur'aini**

**09011182126011**

**Supervisor 1 : Dr. Firdaus, S.T., M.Kom.**

**NIP. 197801212008121003**

**Supervisor 2 : Prof. Dr. Ir. Bambang Tutuko, M.T.**

**NIP. 196001121989031002**

**Approved by,**

**Head of Computer System Department**



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

**196612032006041001**

## HALAMAN PERSETUJUAN

Telah diuji dan lulus pada :

Hari : Jum'at

Tanggal : 23 Mei 2025

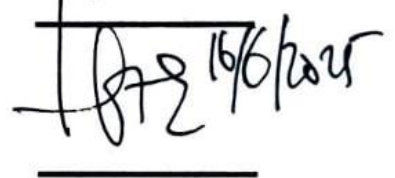
Tim Penguji :

1. Ketua Sidang : Prof. Dr. Erwin, S.Si., M.Si.

2. Penguji Sidang : Dr. M. Fachrurrozi, S.Si., M.T.

3. Pembimbing I : Dr. Firdaus, S.T., M.Kom.

4. Pembimbing II : Prof. Dr. Ir. Bambang Tutuko, M.T.



Mengetahui, 27/6/25

Ketua Jurusan Sistem Komputer



  
Dr. Ir. Sukemi, M.T.

NIP. 196612032006041001

## HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Keisyah Sabinatullah Qur'aini

NIM : 09011182126011

Judul : *Clinical Named Entity Recognition* pada Data Biomedis Menggunakan Model BERT

Hasil Pengecekan Software Turnitin: 5%

Menyatakan bahwa laporan skripsi saya merupakan hasil karya sendiri dan bukan hasil penjiplakan atau plagiat. Apabila ditemukan unsur penjiplakan atau plagiat dalam laporan tugas akhir ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya.

Demikian, pernyataan ini saya buat dalam keadaan sadar dan tanpa paksaan dari siapapun.



Indralaya, 20 Juni 2025

Yang menyatakan,



Keisyah Sabinatullah Qur'aini  
NIM. 09011182126011

## HALAMAN PERSEMBAHAN

1. Rasa syukur yang mendalam saya ucapkan kepada Ayah dan Ibu, yang dengan tulus telah menjadi sumber kekuatan dan cinta tanpa syarat sepanjang hidup saya. Terima kasih atas setiap doa, dukungan, dan semangat yang tak pernah surut, yang telah mengantarkan saya hingga ke tahap ini. Skripsi ini saya dedikasikan sepenuh hati untuk kalian. Ayah, Ibu, putri kecilmu kini telah tumbuh menjadi sosok yang siap melangkah lebih jauh demi menggapai cita-cita.
2. Untuk kakakku tersayang, terima kasih telah menjadi sosok yang selalu hadir dan berarti dalam setiap perjalanan hidupku.

## MOTTO

"Barang siapa menempuh jalan untuk mencari ilmu, maka Allah akan memudahkan baginya jalan menuju surga."

(HR. Muslim)

*"Education is our passport to the future, for tomorrow belongs to the people who prepare for it today."*

(Malcolm X)

"Apa pun yang terjadi, jangan pernah lelah berjuang. Bangkitlah sebagai petarung yang tak mengenal rasa takut, karena setiap lembar buku yang dibaca, setiap malam yang dilalui dengan belajar, adalah langkah kecil menuju impian besar. Kegagalan hanyalah jeda, bukan akhir. Teruslah melangkah, sebab ilmu adalah senjata, dan tekad adalah kunci untuk menaklukkan masa depan."

## KATA PENGANTAR

Assalamu'alaikum Warrahmatullahi Wabarakatuh.

Puji dan syukur atas kehadiran Allah SWT yang telah melimpahkan rahmat, kasih sayang dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi dengan judul “*Clinical Named Entity Recognition* pada Data Biomedis Menggunakan Model BERT”. Shalawat beriring salam senantiasa tercurahkan kepada Nabi Muhammad Shallallaahu'Alaihi Wasallam yang telah membawa kedamaian dan rahmat untuk semesta alam serta menjadi suri tauladan bagi umatnya.

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

1. Allah SWT yang telah melimpahkan berkah serta nikmat kesehatan dan kesempatan kepada penulis dalam mengerjakan skripsi ini.
2. Kedua orang tua, saudara serta keluarga yang telah mendoakan dan memberikan dukungannya baik moril, materiil maupun spiritual kepada penulis hingga dapat menyelesaikan skripsi ini.
3. Bapak Prof. Dr. Erwin, 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 Dr. Firdaus, S.T., M.Kom. dan Prof. Dr. Ir. Bambang Tutuko, M.T. selaku Dosen Pembimbing skripsi yang selalu mengarahkan dan memberi saran serta motivasi kepada saya untuk menyelesaikan skripsi ini.
6. Ibu Prof. Ir. Siti Nurmaini, M.T., Ph.D. selaku *Head of Intelligent System Research Group* (ISysRG) yang telah memberikan kesempatan besar untuk bergabung dan menjadi bagian dari team *research group* ini.
7. Bapak Dr. Ahmad Heryanto, S.Kom., M.T. selaku Dosen Pembimbing Akademik di Jurusan Sistem Komputer Universitas Sriwijaya.

8. Ibu Anggun Islami, M.Kom., Ibu Dr. Ade Iriani Sapitri, M.Kom., Ibu Dr. Annisa Darmawahyuni, M.Kom., Ibu Akhlar Wista Arum, S.T., M.Kom. dan Bapak Naufal Rachmatullah, M.T. sebagai mentor di ISysRG.
9. Pak Angga selaku Administrasi Jurusan Sistem Komputer yang telah membantu dalam proses administrasi terkait skripsi.
10. Shinta Novita Sari selaku sahabat saya dari sekolah dasar sampai sekarang masih setia kebersamaan.
11. Teman-teman seperjuangan di Jurusan Sistem Komputer Angkatan 2021, terutama sahabat saya Indah Gala Putri, Mutiah Andini dan Zahra Hanifa yang selalu memberikan motivasi dan menjadi *support system* saya.
12. Teman-teman satu divisi *text processing* di ISysRG, Muhammad Azriel Apriadi, Tria Lailani dan Tiara Oktarina yang telah menemani dimasa skripsi.

Penulis menyadari bahwa skripsi ini masih jauh dari kata sempurna. Untuk itu penulis sangat terbuka jika ada kritik dan saran yang bersifat membangun agar lebih baik dikemudian hari. Semoga skripsi ini dapat bermanfaat dan dapat menjadi referensi untuk penelitian selanjutnya.

Wassalamu'alaikum Warrahmatullahi Wabarakatuh.

Indralaya, 20 Juni 2025

Penulis,



Keisyah Sabinatullah Qur'aini  
NIM. 09011182126011



# ***CLINICAL NAMED ENTITY RECOGNITION* PADA DATA BIOMEDIS MENGGUNAKAN MODEL BERT**

**KEISYAH SABINATULLAH QUR'AINI (09011182126011)**

Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya

Email : [keisyahsq@gmail.com](mailto:keisyahsq@gmail.com)

## **ABSTRAK**

*Named Entity Recognition* (NER) merupakan salah satu tugas penting dalam pemrosesan bahasa alami (NLP), terutama dalam domain biomedis yang kompleks dan penuh dengan terminologi spesifik. Penelitian ini berfokus pada penerapan model *Clinical Named Entity Recognition* untuk mengekstraksi entitas biomedis dari tiga dataset, yaitu BC2GM, JNLPBA, dan NCBI-*Disease*. Beberapa pendekatan model berbasis BERT digunakan, yakni BERT, BERT yang dikombinasikan dengan BiGRU (BERT-BiGRU), serta BERT yang digabungkan dengan *Support Vector Machine* (BERT-SVM). Evaluasi kinerja dilakukan menggunakan metrik *precision*, *recall*, dan *F1-score* untuk menilai efektivitas dalam mengekstraksi entitas. Hasil eksperimen menunjukkan bahwa model BERT tanpa tambahan arsitektur memberikan performa terbaik pada dua dataset, yaitu BC2GM dan NCBI-*Disease*, dengan *F1-score* masing-masing sebesar 90% dan 92%. Sementara itu, model BERT-BiGRU menunjukkan kinerja terbaik pada dataset JNLPBA dengan *F1-score* sebesar 80%. Temuan ini menunjukkan bahwa meskipun BERT secara umum unggul dalam memahami konteks dan terminologi biomedis, kombinasi BERT dengan BiGRU dapat memberikan keunggulan tambahan dalam kasus tertentu seperti pada dataset JNLPBA.

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

# ***CLINICAL NAMED ENTITY RECOGNITION ON BIOMEDICAL DATA USING BERT MODEL***

**KEISYAH SABINATULLAH QUR'AINI (09011182126011)**

*Computer System Department, Computer Science Faculty, Sriwijaya University*

Email : [keisyahsq@gmail.com](mailto:keisyahsq@gmail.com)

## ***ABSTRACT***

*Named Entity Recognition (NER) is one of the key tasks in natural language processing (NLP), especially within the biomedical domain, which is complex and filled with specific terminologies. This research focuses on the application of a Clinical Named Entity Recognition model to extract biomedical entities from three datasets: BC2GM, JNLPBA, and NCBI-Disease. Several BERT-based model approaches are used, namely BERT, BERT combined with BiGRU (BERT-BiGRU), and BERT combined with Support Vector Machine (BERT-SVM). Performance evaluation is conducted using precision, recall, and F1-score metrics to assess the effectiveness of entity extraction. Experimental results show that the standard BERT model delivers the best performance on two datasets, BC2GM and NCBI-Disease, with F1-scores of 90% and 92%, respectively. Meanwhile, the BERT-BiGRU model achieves the best performance on the JNLPBA dataset, with an F1-score of 80%. These findings indicate that while BERT generally excels in understanding biomedical context and terminology, combining BERT with BiGRU can provide additional advantages in specific cases, such as with the JNLPBA dataset.*

**Keywords :** *Clinical Named Entity Recognition, Biomedical Dataset, Transformer, BERT*

## DAFTAR ISI

|                                                    |              |
|----------------------------------------------------|--------------|
| <b>HALAMAN JUDUL .....</b>                         | <b>i</b>     |
| <b>HALAMAN PENGESAHAN.....</b>                     | <b>ii</b>    |
| <b>AUTHENTICATION PAGE.....</b>                    | <b>iii</b>   |
| <b>HALAMAN PERSETUJUAN.....</b>                    | <b>iv</b>    |
| <b>HALAMAN PERNYATAAN.....</b>                     | <b>v</b>     |
| <b>HALAMAN PERSEMBAHAN.....</b>                    | <b>vi</b>    |
| <b>KATA PENGANTAR.....</b>                         | <b>vii</b>   |
| <b>ABSTRAK .....</b>                               | <b>ix</b>    |
| <b><i>ABSTRACT</i> .....</b>                       | <b>x</b>     |
| <b>DAFTAR ISI.....</b>                             | <b>xi</b>    |
| <b>DAFTAR GAMBAR .....</b>                         | <b>xv</b>    |
| <b>DAFTAR TABEL .....</b>                          | <b>xviii</b> |
| <b>BAB I PENDAHULUAN.....</b>                      | <b>1</b>     |
| 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 .....                     | 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 ..... | 4            |
| 1.6. Sistematika Penulisan.....                    | 5            |
| <b>BAB II TINJAUAN PUSTAKA .....</b>               | <b>6</b>     |
| 2.1. Penelitian Terdahulu .....                    | 6            |
| 2.2. <i>Natural Language Processing</i> .....      | 7            |
| 2.3. <i>Text Pre-processing</i> .....              | 8            |
| 2.3.1. Normalisasi Teks.....                       | 8            |
| 2.3.2. Tokenisasi Teks .....                       | 9            |
| 2.3.3. <i>Word Indexing</i> .....                  | 9            |

|                                                                                                                   |           |
|-------------------------------------------------------------------------------------------------------------------|-----------|
| 2.3.4. <i>Sequence Labelling</i> .....                                                                            | 10        |
| 2.3.5. <i>Padding</i> .....                                                                                       | 11        |
| 2.4. <i>Data Splitting</i> .....                                                                                  | 12        |
| 2.4.1. <i>Stratified Sampling</i> .....                                                                           | 12        |
| 2.4.2. <i>K-Fold Cross Validation</i> .....                                                                       | 13        |
| 2.4.3. <i>Bootstrap Sampling</i> .....                                                                            | 13        |
| 2.5. <i>Dataset Biomedis</i> .....                                                                                | 13        |
| 2.5.1. <i>Dataset BioCreative II Gene Mention</i> .....                                                           | 14        |
| 2.5.2. <i>Dataset Joint Workshop on Natural Language Processing in<br/>Biomedicine and its Applications</i> ..... | 14        |
| 2.5.3. <i>Dataset NCBI-Disease</i> .....                                                                          | 14        |
| 2.6. <i>Clinical Named Entity Recognition</i> .....                                                               | 14        |
| 2.7. <i>Machine Learning</i> .....                                                                                | 15        |
| 2.8. <i>Support Vector Machine</i> .....                                                                          | 15        |
| 2.9. <i>Deep Learning</i> .....                                                                                   | 17        |
| 2.10. <i>Transformer</i> .....                                                                                    | 17        |
| 2.10.1. <i>Transformer's Input and Self-Attention</i> .....                                                       | 18        |
| 2.10.2. <i>Multi-head Mechanism and Normalization</i> .....                                                       | 20        |
| 2.11. <i>Bidirectional Encoder Representation From Transformer</i> .....                                          | 20        |
| 2.12. <i>Gated Recurrent Unit</i> .....                                                                           | 23        |
| 2.13. <i>Pengukuran Kinerja</i> .....                                                                             | 25        |
| 2.13.1. <i>Confusion Matrix</i> .....                                                                             | 25        |
| 2.13.2. <i>Precision</i> .....                                                                                    | 25        |
| 2.13.3. <i>Recall</i> .....                                                                                       | 26        |
| 2.13.4. <i>F1-Score</i> .....                                                                                     | 26        |
| <b>BAB III METODOLOGI PENELITIAN</b> .....                                                                        | <b>27</b> |
| 3.1. <i>Kerangka Kerja</i> .....                                                                                  | 27        |
| 3.2. <i>Akuisisi Data</i> .....                                                                                   | 29        |
| 3.3. <i>Data Splitting</i> .....                                                                                  | 29        |
| 3.4. <i>Exploratory Data Analysis</i> .....                                                                       | 30        |
| 3.4.1. <i>Analisis Deskriptif</i> .....                                                                           | 30        |
| 3.4.2. <i>Visualisasi Perbandingan Panjang Kata dari Setiap Dataset</i> .....                                     | 31        |
| 3.4.3. <i>Visualisasi Persentasi untuk Setiap Label Entitas</i> .....                                             | 32        |
| 3.4.4. <i>Visualisasi Wordcloud untuk Melihat Kata Dominan</i> .....                                              | 33        |

|                                                                                                    |           |
|----------------------------------------------------------------------------------------------------|-----------|
| 3.5. <i>Data Preprocessing</i> .....                                                               | 34        |
| 3.5.1. Normalisasi Teks.....                                                                       | 35        |
| 3.5.2. Tokenisasi Teks .....                                                                       | 36        |
| 3.5.3. <i>Word Indexing</i> .....                                                                  | 37        |
| 3.5.4. <i>Padding</i> .....                                                                        | 38        |
| 3.6. Pelatihan Model.....                                                                          | 38        |
| 3.6.1. Model BERT .....                                                                            | 39        |
| 3.6.2. Model BERT-BiGRU .....                                                                      | 44        |
| 3.7. Penambahan Model BERT-SVM sebagai Komparasi .....                                             | 46        |
| 3.8. Evaluasi Model.....                                                                           | 48        |
| <b>BAB IV HASIL DAN ANALISIS.....</b>                                                              | <b>49</b> |
| 4.1. Skenario Percobaan Awal .....                                                                 | 49        |
| 4.2. Hasil Percobaan pada Dataset BC2GM.....                                                       | 50        |
| 4.3. Hasil Percobaan pada Dataset JNLPBA .....                                                     | 56        |
| 4.4. Hasil Percobaan pada Dataset NCBI- <i>Disease</i> .....                                       | 64        |
| 4.5. <i>Learning Curve</i> pada Data Latih dan Validasi untuk Mengevaluasi<br>Performa model ..... | 70        |
| 4.6. <i>Macro Average</i> Dataset BC2GM, JNLPBA dan NCBI- <i>Disease</i> .....                     | 73        |
| 4.6.1. <i>Macro Average</i> Ketiga Dataset Biomedis dengan <i>Batch Size</i> 16.....               | 73        |
| 4.6.2. <i>Macro Average</i> Ketiga Dataset Biomedis dengan <i>Batch Size</i> 32.....               | 79        |
| 4.7. <i>Confusion Matrix</i> pada Data Uji.....                                                    | 86        |
| 4.7.1. <i>Confusion Matrix</i> pada Dataset BC2GM.....                                             | 86        |
| 4.7.2. <i>Confusion Matrix</i> pada Dataset JNLPBA.....                                            | 88        |
| 4.7.3. <i>Confusion matrix</i> pada Dataset NCBI- <i>Disease</i> .....                             | 91        |
| 4.8. Durasi Waktu <i>Training</i> Model BERT dan BERT-BiGRU .....                                  | 93        |
| 4.9. Peningkatan Metrik Evaluasi Model pada 3 Dataset biomedis .....                               | 95        |
| 4.9.1. Skenario Percobaan Akhir .....                                                              | 95        |
| 4.9.2. Hasil Percobaan pada Dataset BC2GM .....                                                    | 96        |
| 4.9.3. Hasil Percobaan pada Dataset JNLPBA .....                                                   | 102       |
| 4.9.4. Hasil Percobaan pada Dataset NCBI- <i>Disease</i> .....                                     | 109       |
| 4.9.5. <i>Macro Average</i> pada Dataset BC2GM, JNLPBA dan NCBI- <i>Disease</i> .....              | 115       |
| 4.9.6. <i>Confusion Matrix</i> pada Data Uji BC2GM, JNLPBA dan NCBI- <i>Disease</i> .....          | 123       |
| 4.9.7. Rangkuman Waktu <i>Training</i> pada Ketiga Model .....                                     | 129       |

|                                                                    |            |
|--------------------------------------------------------------------|------------|
| 4.9.8. Rangkuman Perbandingan Hasil Percobaan Awal dan Akhir ..... | 129        |
| 4.9.9. Komparasi Hasil Penelitian dengan Penelitian Terdahulu..... | 132        |
| <b>BAB V KESIMPULAN DAN SARAN .....</b>                            | <b>134</b> |
| 5.1. Kesimpulan.....                                               | 134        |
| 5.2. Saran .....                                                   | 135        |
| <b>DAFTAR PUSTAKA.....</b>                                         | <b>136</b> |
| <b>LAMPIRAN.....</b>                                               | <b>142</b> |

## DAFTAR GAMBAR

|                                                                                                                                   |    |
|-----------------------------------------------------------------------------------------------------------------------------------|----|
| Gambar 2.1 Ilustrasi Normalisasi Teks .....                                                                                       | 9  |
| Gambar 2.2 Ilustrasi Tokenisasi Teks.....                                                                                         | 9  |
| Gambar 2.3 <i>Sequence Labelling</i> .....                                                                                        | 10 |
| Gambar 2.4 Ilustrasi <i>Padding</i> .....                                                                                         | 11 |
| Gambar 2.5 Ilustrasi <i>Data Splitting</i> .....                                                                                  | 12 |
| Gambar 2.6 <i>Hyperplane</i> Model SVM.....                                                                                       | 16 |
| Gambar 2.7 Arsitektur <i>Transformer</i> .....                                                                                    | 18 |
| Gambar 2.8 Arsitektur BERT .....                                                                                                  | 21 |
| Gambar 2.9 Skema Alur Model BERT .....                                                                                            | 22 |
| Gambar 2.10 Arsitektur GRU .....                                                                                                  | 24 |
| Gambar 2.11 <i>Layer</i> BiGRU .....                                                                                              | 24 |
| Gambar 3.1 Rancangan Kerangka Kerja.....                                                                                          | 28 |
| Gambar 3.2 Sampel Teks Dataset. ....                                                                                              | 30 |
| Gambar 3.3 Perbandingan Panjang Kata Setiap Dataset. ....                                                                         | 32 |
| Gambar 3.4 Perbandingan Persentase Label Entitas pada a) Dataset BC2GM, b) Dataset JNLPBA, c) Dataset NCBI- <i>Disease</i> . .... | 33 |
| Gambar 3.5 Frekuensi Kata Dominan pada 3 Dataset Biomedis.....                                                                    | 34 |
| Gambar 3.6 <i>Flowchart Data Preprocessing</i> . ....                                                                             | 35 |
| Gambar 3.7 Ilustrasi Normalisasi Teks. ....                                                                                       | 36 |
| Gambar 3.8 <i>Wordpiece Tokenization</i> . ....                                                                                   | 37 |
| Gambar 3.9 Ilustrasi <i>Word Indexing</i> .....                                                                                   | 37 |
| Gambar 3.10 Ilustrasi <i>Padding</i> . ....                                                                                       | 38 |
| Gambar 3.11 Arsitektur BERT. ....                                                                                                 | 39 |
| Gambar 3.12 <i>Input</i> Token pada Model BERT.....                                                                               | 39 |
| Gambar 3.13 <i>Input</i> IDs.....                                                                                                 | 40 |
| Gambar 3.14 <i>Input</i> IDs pada <i>Layer Input</i> Model BERT. ....                                                             | 41 |
| Gambar 3.15 <i>Embedding Layer</i> pada Model BERT.....                                                                           | 41 |
| Gambar 3.16 Ilustrasi Proses pada <i>Embedding Layer</i> Model BERT. ....                                                         | 42 |
| Gambar 3.17 <i>Hidden State</i> dan Prediksi Probabilitas <i>Output</i> . ....                                                    | 43 |
| Gambar 3.18 Arsitektur Model BERT-BiGRU. ....                                                                                     | 44 |

|                                                                                                       |     |
|-------------------------------------------------------------------------------------------------------|-----|
| Gambar 3.19 Arsitektur Model BERT-SVM.....                                                            | 46  |
| Gambar 3.20 Vektor <i>Embedding</i> BERT Diproses Dalam SVM.....                                      | 47  |
| Gambar 4.1 <i>Heatmap</i> Dataset BC2GM dengan Metrik <i>Precision</i> .....                          | 51  |
| Gambar 4.2 <i>Heatmap</i> Dataset BC2GM dengan Metrik <i>Recall</i> .....                             | 53  |
| Gambar 4.3 <i>Heatmap</i> Dataset BC2GM dengan Metrik <i>F1-Score</i> . ....                          | 55  |
| Gambar 4.4 <i>Heatmap</i> Dataset JNLPBA dengan Metrik <i>Precision</i> . ....                        | 58  |
| Gambar 4.5 <i>Heatmap</i> Dataset JNLPBA dengan Metrik <i>Recall</i> . ....                           | 61  |
| Gambar 4.6 <i>Heatmap</i> Dataset JNLPBA dengan Metrik <i>F1-Score</i> .....                          | 63  |
| Gambar 4.7 <i>Heatmap</i> Dataset NCBI-Disease dengan Metrik <i>Precision</i> . ....                  | 65  |
| Gambar 4.8 <i>Heatmap</i> Dataset NCBI-Disease dengan Metrik <i>Recall</i> .....                      | 67  |
| Gambar 4.9 <i>Heatmap</i> Dataset NCBI-Disease dengan Metrik <i>F1-Score</i> .....                    | 69  |
| Gambar 4.10 <i>Learning Curve</i> Terbaik pada Ketiga Dataset Biomedis.....                           | 72  |
| Gambar 4.11 <i>Bar Chart Macro Average</i> Metrik <i>Precision</i> pada Ketiga Dataset Biomedis. .... | 75  |
| Gambar 4.12 <i>Bar Chart Macro Average</i> Metrik <i>Recall</i> pada Ketiga Dataset Biomedis. ....    | 77  |
| Gambar 4.13 <i>Bar Chart Macro Average</i> Metrik <i>F1-Score</i> pada Ketiga Dataset Biomedis. ....  | 79  |
| Gambar 4.14 <i>Bar Chart Macro Average</i> Metrik <i>Precision</i> pada Ketiga Dataset Biomedis. .... | 81  |
| Gambar 4.15 <i>Bar Chart Macro Average</i> Metrik <i>Recall</i> pada Ketiga Dataset Biomedis. ....    | 83  |
| Gambar 4.16 <i>Bar Chart Macro Average</i> Metrik <i>F1-Score</i> pada Ketiga Dataset Biomedis. ....  | 85  |
| Gambar 4.17 <i>Confusion Matrix</i> pada Dataset BC2GM. ....                                          | 88  |
| Gambar 4.18 <i>Confusion Matrix</i> pada Dataset JNLPBA. ....                                         | 90  |
| Gambar 4.19 <i>Confusion Matrix</i> Dataset NCBI-Disease. ....                                        | 93  |
| Gambar 4.20 <i>Heatmap</i> Dataset BC2GM dengan Metrik <i>Precision</i> .....                         | 97  |
| Gambar 4.21 <i>Heatmap</i> Dataset BC2GM dengan Metrik <i>Recall</i> .....                            | 99  |
| Gambar 4.22 <i>Heatmap</i> Dataset BC2GM dengan Metrik <i>F1-Score</i> . ....                         | 101 |
| Gambar 4.23 <i>Heatmap</i> Dataset JNLPBA dengan Metrik <i>Precision</i> . ....                       | 103 |
| Gambar 4.24 <i>Heatmap</i> Dataset JNLPBA dengan Metrik <i>Recall</i> . ....                          | 106 |



|                                                                                                                                             |     |
|---------------------------------------------------------------------------------------------------------------------------------------------|-----|
| Gambar 4.25 <i>Heatmap</i> Dataset JNLPBA dengan Metrik F1-Score.....                                                                       | 108 |
| Gambar 4.26 <i>Heatmap</i> Dataset NCBI-Disease dengan Metrik <i>Precision</i> .....                                                        | 110 |
| Gambar 4.27 <i>Heatmap</i> Dataset NCBI-Disease dengan Metrik <i>Recall</i> .....                                                           | 112 |
| Gambar 4.28 <i>Heatmap</i> Dataset NCBI-Disease dengan Metrik F1-Score.....                                                                 | 114 |
| Gambar 4.29 <i>Bar Chart Macro Average</i> Metrik <i>Precision</i> pada Ketiga Dataset Biomedis. ....                                       | 117 |
| Gambar 4.30 <i>Bar Chart Macro Average</i> Metrik <i>Recall</i> pada Ketiga Dataset Biomedis. ....                                          | 119 |
| Gambar 4.31 <i>Bar Chart Macro Average</i> Metrik F1-Score pada Ketiga Dataset Biomedis. ....                                               | 121 |
| Gambar 4.32 <i>Confusion Matrix</i> Dataset BC2GM.....                                                                                      | 124 |
| Gambar 4.33 <i>Confusion Matrix</i> Dataset JNLPBA.....                                                                                     | 126 |
| Gambar 4.34 <i>Confusion Matrix</i> Dataset NCBI-Disease.....                                                                               | 128 |
| Gambar 4.35 <i>Macro Average</i> F1-Score Terbaik dari Percobaan Awal dan Percobaan Akhir pada Dataset BC2GM, JNLPBA dan NCBI-Disease. .... | 130 |

## DAFTAR TABEL

|                                                                                                      |     |
|------------------------------------------------------------------------------------------------------|-----|
| Tabel 2.1 <i>Confusion Matrix</i> .....                                                              | 25  |
| Tabel 3.1 Persentase dan Jumlah Data Latih, Data Validasi dan Data Uji Ketiga Dataset Biomedis ..... | 29  |
| Tabel 3.2 Keterangan Label Entitas Dataset Biomedis .....                                            | 31  |
| Tabel 3.3 <i>Hyperparameter</i> pada Penelitian.....                                                 | 45  |
| Tabel 3.4 <i>Hyperparameter</i> Model BERT-SVM .....                                                 | 47  |
| Tabel 4.1 Skenario Percobaan .....                                                                   | 49  |
| Tabel 4.2 Hasil Percobaan Dataset BC2GM dengan Metrik <i>Precision</i> .....                         | 50  |
| Tabel 4.3 Hasil Percobaan Dataset BC2GM dengan Metrik <i>Recall</i> . ....                           | 52  |
| Tabel 4.4 Hasil Percobaan Dataset BC2GM dengan Metrik <i>F1-Score</i> .....                          | 54  |
| Tabel 4.5 Hasil Percobaan Dataset JNLPBA dengan Metrik <i>Precision</i> .....                        | 56  |
| Tabel 4.6 Hasil Percobaan Dataset JNLPBA dengan Metrik <i>Recall</i> .....                           | 59  |
| Tabel 4.7 Hasil Percobaan Dataset JNLPBA dengan Metrik <i>F1-Score</i> .....                         | 61  |
| Tabel 4.8 Hasil Percobaan pada Dataset NCBI- <i>Disease</i> dengan Metrik <i>Precision</i> .....     | 64  |
| Tabel 4.9 Hasil Percobaan pada Dataset NCBI- <i>Disease</i> dengan Metrik <i>Recall</i> ..           | 66  |
| Tabel 4.10 Hasil Percobaan pada Dataset NCBI- <i>Disease</i> dengan Metrik <i>F1-Score</i> .....     | 68  |
| Tabel 4.11 <i>Macro Average</i> Metrik <i>Precision</i> .....                                        | 73  |
| Tabel 4.12 <i>Macro Average</i> Metrik <i>Recall</i> .....                                           | 75  |
| Tabel 4.13 <i>Macro average</i> Metrik <i>F1-Score</i> .....                                         | 77  |
| Tabel 4.14 <i>Macro Average</i> Metrik <i>Precision</i> .....                                        | 79  |
| Tabel 4.15 <i>Macro Average</i> Metrik <i>Recall</i> .....                                           | 81  |
| Tabel 4.16 <i>Macro Average</i> Metrik <i>F1-Score</i> .....                                         | 83  |
| Tabel 4.17 Durasi Waktu <i>Training</i> Model BERT dan BERT-BiGRU .....                              | 94  |
| Tabel 4.18 Skenario Percobaan Akhir .....                                                            | 95  |
| Tabel 4.19 Hasil Percobaan Dataset BC2GM dengan Metrik <i>Precision</i> .....                        | 96  |
| Tabel 4.20 Hasil Percobaan Dataset BC2GM dengan Metrik <i>Recall</i> .....                           | 98  |
| Tabel 4.21 Hasil percobaan dataset BC2GM dengan metrik <i>F1-Score</i> .....                         | 100 |
| Tabel 4.22 Hasil Percobaan pada Dataset JNLPBA dengan Metrik <i>Precision</i> ...                    | 102 |

|                                                                                                      |     |
|------------------------------------------------------------------------------------------------------|-----|
| Tabel 4.23 Hasil Percobaan pada Dataset JNLPBA dengan Metrik <i>Recall</i> .....                     | 104 |
| Tabel 4.24 Hasil Percobaan pada Dataset JNLPBA dengan Metrik <i>F1-Score</i> ....                    | 106 |
| Tabel 4.25 Hasil Percobaan pada Dataset NCBI- <i>Disease</i> dengan Metrik <i>Precision</i><br>..... | 109 |
| Tabel 4.26 Hasil Percobaan pada Dataset NCBI- <i>Disease</i> dengan Metrik <i>Recall</i>             | 111 |
| Tabel 4.27 Hasil Percobaan pada Dataset NCBI- <i>Disease</i> dengan Metrik <i>F1-Score</i><br>.....  | 113 |
| Tabel 4.28 <i>Macro Average</i> dengan Metrik <i>Precision</i> .....                                 | 116 |
| Tabel 4.29 <i>Macro Average</i> dengan Metrik <i>Recall</i> .....                                    | 118 |
| Tabel 4.30 <i>Macro Average</i> dengan Metrik <i>F1-Score</i> .....                                  | 120 |
| Tabel 4.31 Rangkuman Waktu <i>Training</i> pada Model BERT, BERT-BiGRU dan<br>BERT-SVM .....         | 129 |
| Tabel 4.32 Komparasi Model BERT pada Dataset Umum dan Dataset Biomedis<br>.....                      | 132 |
| Tabel 4.33 Komparasi Semua Percobaan dengan Penelitian Terdahulu pada Dataset<br>Biomedis .....      | 132 |

# BAB I

## PENDAHULUAN

### 1.1. Latar Belakang

Perkembangan teknologi di bidang *Natural Language Processing* (NLP) telah membawa kemajuan signifikan dalam tugas *Named Entity Recognition* (NER), baik pada data umum maupun data khusus seperti data medis [1][2]. NER merupakan tugas penting dalam NLP yang bertujuan untuk mengidentifikasi dan mengklasifikasikan entitas tertentu dari teks tidak terstruktur ke dalam kategori yang telah ditentukan, seperti nama orang, lokasi, atau istilah medis [3][4]. Pada data umum, berbagai pendekatan NER telah berhasil diterapkan, terutama sejak munculnya model berbasis *transformer* seperti *Bidirectional Encoder Representations from Transformers* (BERT) yang mampu memahami konteks kata secara lebih mendalam melalui mekanisme *self-attention* [5][6].

Namun, ketika pendekatan yang berhasil pada data umum ini diterapkan ke domain khusus seperti data medis atau biomedis, muncul tantangan baru. Teks medis memiliki karakteristik yang berbeda dengan teks umum, seperti penggunaan istilah teknis, singkatan medis, serta struktur kalimat yang tidak standar [7]. Oleh karena itu, penerapan model NER yang efektif pada data medis memerlukan penyesuaian, baik dari sisi arsitektur model maupun data pelatihan yang digunakan.

Beberapa penelitian sebelumnya telah mencoba mengadaptasi model NER umum seperti BERT untuk tugas *Clinical Named Entity Recognition* (CNER), yaitu tugas NER khusus pada teks klinis dan biomedis [8]. Hasil penelitian sebelumnya bervariasi tergantung jenis data dan arsitektur yang digunakan. Oleh karena itu, evaluasi performa model NER pada data umum dan biomedis penting untuk mengetahui sejauh mana BERT dapat dioptimalkan dalam domain medis.

Penelitian ini bertujuan untuk melakukan percobaan penerapan model NER berbasis *transformer*, yaitu BERT, dalam tugas CNER menggunakan beberapa dataset biomedis. Selain itu, penelitian ini juga mengkaji pengaruh kombinasi arsitektur seperti BERT-BiGRU serta integrasi algoritma *machine learning* seperti *Support Vector Machine* (SVM) terhadap performa ekstraksi entitas medis. BiGRU merupakan versi *bidirectional* dari GRU, salah satu varian dari *Recurrent Neural*

*Network* (RNN), yang memproses data sekuensial dari dua arah yaitu maju (*forward*) dan mundur (*backward*) sehingga mampu menangkap konteks informasi dari kedua sisi dalam suatu urutan [9]. Sementara itu, model SVM merupakan pengklasifikasi yang sangat berguna untuk data yang tidak terdistribusi secara merata atau tidak beraturan, dan dapat diterapkan pada berbagai jenis data seperti teks, gambar, audio, maupun jenis data lainnya [10].

Dengan melakukan perbandingan terhadap pendekatan-pendekatan sebelumnya baik pada data umum maupun medis, penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan model CNER yang lebih efektif untuk domain biomedis.

## 1.2. Rumusan Masalah

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

1. Bagaimana menerapkan CNER dengan model BERT yang telah dilatih pada data teks umum agar secara akurat mengenali entitas pada data biomedis?
2. Bagaimana performa model BERT yang telah dilatih pada data teks umum dalam mengenali entitas pada data biomedis pada kasus CNER?
3. Bagaimana pengaruh kombinasi model BERT dan model BiGRU dalam mengenali entitas pada data biomedis untuk tugas CNER?
4. Bagaimana pengaruh kombinasi model BERT dan model SVM dalam mengenali entitas pada data biomedis untuk tugas CNER?

## 1.3. Batasan Masalah

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

1. Penelitian ini menggunakan metode *deep learning* berbasis *transformer* untuk melakukan CNER.
2. Penelitian ini terbatas pada penggunaan tiga dataset yang bertujuan untuk ekstraksi informasi biomedis dan anotasi teks. Dataset yang digunakan dalam penelitian ini adalah BC2GM, JNLPBA, dan NCBI-*Disease*.

3. Model yang digunakan untuk melatih ketiga dataset tersebut adalah BERT, yang memanfaatkan 12 lapisan pada arsitektur *transformer*.
4. Penelitian ini membatasi evaluasi kinerja model CNER hanya menggunakan tiga metrik utama, yaitu *F1-score*, *precision*, dan *recall*, karena lebih efektif dalam menilai kinerja NER, menyeimbangkan kesalahan, dan lebih akurat dibandingkan *accuracy*.

#### 1.4. Tujuan

Tujuan yang akan dicapai dalam penelitian ini adalah sebagai berikut:

1. Menerapkan sebuah model *deep learning* berbasis *transformer*, model BERT yang digunakan untuk CNER yang telah dilatih dengan data teks umum kedalam data biomedis.
2. Mengevaluasi performa model BERT yang telah dilatih pada data teks umum dalam mengenali entitas pada data biomedis untuk tugas CNER menggunakan berbagai metrik evaluasi.
3. Menganalisis pengaruh kombinasi model BERT dan model BiGRU dalam mengenali entitas pada data biomedis untuk tugas CNER.
4. Menganalisis pengaruh kombinasi model BERT dan model SVM dalam mengenali entitas pada data biomedis untuk tugas CNER.

#### 1.5. Metodologi Penulisan

Metode yang digunakan dalam penulisan tugas akhir ini, mulai dari studi pustaka hingga penarikan kesimpulan dan saran. Setiap langkah metode dijelaskan secara rinci untuk memberikan pemahaman yang menyeluruh 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 literatur yang terdapat pada buku, *paper* dan internet mengenai “*Clinical Named Entity Recognition Pada Data Biomedis Menggunakan Model BERT*”. Metode ini bertujuan untuk memperoleh dasar teori yang relevan dan memahami pendekatan penelitian sebelumnya sebagai acuan.

### 1.5.2. Metode Konsultasi

Metode ini dilakukan dengan konsultasi dengan ahli atau pihak yang memiliki pengetahuan dan pengalaman dalam menangani permasalahan yang dihadapi dalam penulisan tugas akhir.

### 1.5.3. Metode Pembuatan Model

Metode ini mencakup proses perancangan dan pembangunan model yang digunakan dalam penelitian. Pemodelan dilakukan melalui serangkaian eksperimen dengan menggunakan pendekatan *deep learning* berbasis *transformer*, yaitu BERT, serta kombinasi arsitektur BERT-BiGRU dan integrasi dengan algoritma *machine learning* seperti SVM. Setiap model diterapkan dan diuji pada beberapa dataset biomedis untuk mengevaluasi kinerjanya dalam tugas CNER.

### 1.5.4. Metode Pengujian dan Validasi

Metode ini melakukan pengujian terhadap CNER dengan model yang telah dibuat untuk menentukan apakah model tersebut dapat menghasilkan nilai yang baik atau tidak. Evaluasi dilakukan untuk menilai efektivitas model dalam mengekstraksi entitas medis berdasarkan metrik seperti *precision*, *recall*, dan *F1-score*.

### 1.5.5. Metode Hasil dan Analisa

Hasil dari pengujian pada penelitian ini akan dianalisis kelebihan dan kekurangan dari model CNER yang telah dibuat, sehingga dapat digunakan dan diimprovisasi untuk penelitian selanjutnya. Analisis ini bertujuan mengevaluasi efektivitas model dalam mengekstraksi entitas medis serta mengidentifikasi aspek yang perlu ditingkatkan.

### 1.5.6. Metode Penarikan Kesimpulan dan Saran

Pada tahap ini, peneliti menyimpulkan hasil evaluasi performa model CNER yang telah diterapkan pada data biomedis, serta memberikan saran yang dapat menjadi acuan untuk pengembangan dan penelitian lebih lanjut di bidang ekstraksi entitas medis.

## **1.6. Sistematika Penulisan**

Sistematika yang digunakan dalam penulisan tugas akhir ini adalah:

### **BAB I PENDAHULUAN**

Bab I memberikan uraian tentang awal dari suatu penulisan, meliputi latar belakang, perumusan dan batasan masalah, tujuan dan manfaat, metodologi penelitian, serta sistematika penulisan.

### **BAB II TINJAUAN PUSTAKA**

Bab II memaparkan mengenai teori – teori dasar yang menjadi landasan dari penelitian yang dilakukan.

### **BAB III METODOLOGI PENELITIAN**

Bab III berisi penjelasan detail mengenai teknik, metode, serta alur proses yang digunakan dalam penelitian.

### **BAB IV HASIL DAN ANALISIS**

Bab IV menjelaskan hasil pengujian yang diperoleh dan menjelaskan analisa terhadap hasil penelitian yang telah dilakukan.

### **BAB V KESIMPULAN DAN SARAN**

Bab V berisi kesimpulan dari hasil dan analisa dari keseluruhan penelitian yang dilakukan dan saran untuk penelitian selanjutnya.

### **DAFTAR PUSTAKA**

Daftar pustaka berisi daftar referensi 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] O. Solarte-Pabón *et al.*, “Transformers for extracting breast cancer information from Spanish clinical narratives,” *Artif. Intell. Med.*, vol. 143, no. May, p. 102625, 2023, doi: 10.1016/j.artmed.2023.102625.
- [2] S. Silalahi, T. Ahmad, and H. Studiawan, “Transformer-Based Named Entity Recognition on Drone Flight Logs to Support Forensic Investigation,” *IEEE Access*, vol. 11, no. January, pp. 3257–3274, 2023, doi: 10.1109/ACCESS.2023.3234605.
- [3] Q. Zhang, Y. Sun, L. Zhang, Y. Jiao, and Y. Tian, “Named entity recognition method in health preserving field based on BERT,” *Procedia Comput. Sci.*, vol. 183, no. 2018, pp. 212–220, 2021, doi: 10.1016/j.procs.2021.03.010.
- [4] C. Dou, X. Sun, Y. Wang, Y. Ji, B. Ma, and X. Li, “Domain-Adapted Dependency Parsing for Cross-Domain Named Entity Recognition,” *Proc. 37th AAAI Conf. Artif. Intell. AAAI 2023*, vol. 37, pp. 12737–12744, 2023, doi: 10.1609/aaai.v37i11.26498.
- [5] V. K. Prasad, “Customized Name Entity Recognition for Medical Data using BERT,” vol. 11, no. 07, pp. 330–332, 2022.
- [6] D. H. Fudholi, A. Zahra, S. Rani, S. N. Huda, I. V. Paputungan, and Z. Zukhri, “BERT-based tourism named entity recognition: making use of social media for travel recommendations,” *PeerJ Comput. Sci.*, vol. 9, 2023, doi: 10.7717/PEERJ-CS.1731.
- [7] W. Lu *et al.*, “Application of Entity-BERT model based on neuroscience and brain-like cognition in electronic medical record entity recognition,” *Front. Neurosci.*, vol. 17, 2023, doi: 10.3389/fnins.2023.1259652.
- [8] Y. Wu, J. Huang, C. Xu, H. Zheng, L. Zhang, and J. Wan, “Research on Named Entity Recognition of Electronic Medical Records Based on RoBERTa and Radical-Level Feature,” *Wirel. Commun. Mob. Comput.*, vol. 2021, 2021, doi: 10.1155/2021/2489754.

- [9] X. Lv *et al.*, “Chinese Named Entity Recognition in the Geoscience Domain Based on BERT,” *Earth Sp. Sci.*, vol. 9, no. 3, pp. 1–15, 2022, doi: 10.1029/2021EA002166.
- [10] A. T. Imam, A. Alhroob, and W. J. Alzyadat, “SVM Machine Learning Classifier to Automate the Extraction of SRS Elements,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 3, pp. 174–185, 2021, doi: 10.14569/IJACSA.2021.0120322.
- [11] M. M. Nczuk, “TRANSFORMER-BASED NAMED ENTITY RECOGNITION WITH COMBINED DATA REPRESENTATION,” no. 2019, 2021.
- [12] M. C. Cariello, A. Lenci, and R. Mitkov, “A Comparison between Named Entity Recognition Models in the Biomedical Domain,” pp. 76–84, 2022, doi: 10.26615/978-954-452-071-7\_009.
- [13] 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.
- [14] M. B. Shishehgarkhaneh, R. C. Moehler, Y. Fang, A. A. Hijazi, and H. Aboutorab, “Transformer-Based Named Entity Recognition in Construction Supply Chain Risk Management in Australia,” *IEEE Access*, vol. 12, no. March, pp. 41829–41851, 2024, doi: 10.1109/ACCESS.2024.3377232.
- [15] Y. M. Kim and T. H. Lee, “Korean clinical entity recognition from diagnosis text using BERT,” *BMC Med. Inform. Decis. Mak.*, vol. 20, no. Suppl 7, pp. 1–9, 2020, doi: 10.1186/s12911-020-01241-8.
- [16] S. B. Panuntun, D. Krismawati, S. Pramana, and E. T. Astuti, “Analisis Teks Pemberitaan Telemedicine di Indonesia: Pendekatan Sentimen, NER, Topic Modeling, dan Social Network dalam Memahami Isu dan Persepsi,” *Indones. Heal. Inf. Manag. J.*, vol. 11, no. 1, pp. 56–67, 2023, doi: 10.47007/inohim.v11i1.500.

- [17] A. Javaloy and G. García-Mateos, “Text normalization using encoder-decoder networks based on the causal feature extractor,” *Appl. Sci.*, vol. 10, no. 13, 2020, doi: 10.3390/app10134551.
- [18] J. E. Br Sinulingga and H. C. K. Sitorus, “Analisis Sentimen Opini Masyarakat terhadap Film Horor Indonesia Menggunakan Metode SVM dan TF-IDF,” *J. Manaj. Inform.*, vol. 14, no. 1, pp. 42–53, 2024, doi: 10.34010/jamika.v14i1.11946.
- [19] N. Hafidz and D. Yanti Liliana, “Klasifikasi Sentimen pada Twitter Terhadap WHO Terkait Covid-19 Menggunakan SVM, N-Gram, PSO,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 213–219, 2021, doi: 10.29207/resti.v5i2.2960.
- [20] S. Liu, “Task-Adaptive Tokenization: Enhancing Long-Form Text Generation Efficacy in Mental Health and Beyond,” pp. 15264–15281, 2023.
- [21] K. Dalimunthe and B. H. Hayadi, “Information Text Retrieval Untuk Pencarian Data Penilaian Mengacu Pada Saran Dari Pengunjung Menggunakan Vector Space Modelimplementasi,” *J. Comput. Sci. Inf. Technol. Progr. Stud. Teknol. Inf.*, vol. 3, no. 2, pp. 73–79, 2022, [Online]. Available: <http://jurnal.ulb.ac.id/index.php/JCoInT/index>
- [22] Z. He, Z. Wang, W. Wei, S. Feng, X. Mao, and S. Jiang, “A Survey on Recent Advances in Sequence Labeling from Deep Learning Models,” pp. 1–16, 2020, [Online]. Available: <http://arxiv.org/abs/2011.06727>
- [23] T. Kato, K. Abe, H. Ouchi, S. Miyawaki, J. Suzuki, and K. Inui, “Embeddings of label components for sequence labeling: A case study of fine-grained named entity recognition,” *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 222–229, 2020, doi: 10.18653/v1/2020.acl-srw.30.
- [24] A. Asroni, K. R. Ku-Mahamud, C. Damarjati, and H. B. Slamati, “Arabic speech classification method based on padding and deep learning neural network,” *Baghdad Sci. J.*, vol. 18, no. 2, pp. 925–936, 2021, doi: 10.21123/bsj.2021.18.2(Suppl.).0925.

- [25] A. Lopez-del Rio, M. Martin, A. Perera-Lluna, and R. Saidi, "Effect of sequence padding on the performance of deep learning models in archaeal protein functional prediction," *Sci. Rep.*, vol. 10, no. 1, pp. 1–14, 2020, doi: 10.1038/s41598-020-71450-8.
- [26] D. E. Birba, "A Comparative study of data splitting algorithms for machine learning model selection," *Degree Proj. Comput. Sci. Eng.*, vol. 2020, no. 1, pp. 1–23, 2020, [Online]. Available: <https://www.diva-portal.org/smash/get/diva2:1506870/FULLTEXT01.pdf>
- [27] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Min.*, vol. 15, no. 4, pp. 531–538, 2022, doi: 10.1002/sam.11583.
- [28] J. Sadaiyandi, P. Arumugam, A. K. Sangaiah, and C. Zhang, "Stratified Sampling-Based Deep Learning Approach to Increase Prediction Accuracy of Unbalanced Dataset," *Electron.*, vol. 12, no. 21, pp. 1–16, 2023, doi: 10.3390/electronics12214423.
- [29] U. Michelucci and F. Venturini, "Estimating Neural Network's Performance with Bootstrap: A Tutorial," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 2, pp. 357–373, 2021, doi: 10.3390/make3020018.
- [30] U. Naseem, K. Musial, P. Eklund, and M. Prasad, "Biomedical Named-Entity Recognition by Hierarchically Fusing BioBERT Representations and Deep Contextual-Level Word-Embedding," *Proc. Int. Jt. Conf. Neural Networks*, 2020, doi: 10.1109/IJCNN48605.2020.9206808.
- [31] N. Alsaaran and M. Alrabiah, "Classical Arabic Named Entity Recognition Using Variant Deep Neural Network Architectures and BERT," *IEEE Access*, vol. 9, pp. 91537–91547, 2021, doi: 10.1109/ACCESS.2021.3092261.
- [32] C. Chazar and B. Erawan, "Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine," *Inf. (Jurnal Inform. dan Sist. Informatika)*, vol. 12, no. 1, pp. 67–80, 2020, doi: 10.37424/informatika.v12i1.48.

- [33] S. U. Hassan, J. Ahamed, and K. Ahmad, “Analytics of machine learning-based algorithms for text classification,” *Sustain. Oper. Comput.*, vol. 3, no. April, pp. 238–248, 2022, doi: 10.1016/j.susoc.2022.03.001.
- [34] R. B. Afandi, T. F. Nurdiansyah, A. N. Ramadhani, and A. P. Sari, “Implementasi Support Vector Machine untuk Analisis Sentimen Aplikasi ‘Mpstore - Super App UMKM,’” *J. Inform. Polinema*, vol. 10, no. 4, pp. 565–570, 2024, doi: 10.33795/jip.v10i4.5427.
- [35] P. Chen, M. Zhang, X. Yu, and S. Li, “Named entity recognition of Chinese electronic medical records based on a hybrid neural network and medical MC-BERT,” *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 1–13, 2022, doi: 10.1186/s12911-022-02059-2.
- [36] T. Xiao and J. Zhu, “Introduction to Transformers: an NLP Perspective,” 2023, [Online]. Available: <http://arxiv.org/abs/2311.17633>
- [37] N. M. Foumani, C. W. Tan, G. I. Webb, and M. Salehi, “Improving position encoding of transformers for multivariate time series classification,” *Data Min. Knowl. Discov.*, vol. 38, no. 1, pp. 22–48, 2024, doi: 10.1007/s10618-023-00948-2.
- [38] Q. Qin, S. Zhao, and C. Liu, “A BERT-BiGRU-CRF Model for Entity Recognition of Chinese Electronic Medical Records,” *Complexity*, vol. 2021, no. 1, 2021, doi: 10.1155/2021/6631837.
- [39] R. Ma, H. Cao, Z. Song, and X. Wu, “Study on Chinese Semantic Entity Recognition Method for Cabin Utilizing BERT-BiGRU Model,” *IEEE Access*, vol. 12, no. March, pp. 56042–56049, 2024, doi: 10.1109/ACCESS.2024.3386760.
- [40] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, “Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking,” *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.
- [41] R. Mas, R. W. Panca, K. Atmaja1, and W. Yustanti2, “Analisis Sentimen

Customer Review Aplikasi Ruang Guru dengan Metode BERT (Bidirectional Encoder Representations from Transformers),” *Jeisbi*, vol. 02, no. 03, p. 2021, 2021.

- [42] A. R. Hanum *et al.*, “MENDETEKSI BERITA HOAKS PERFORMANCE ANALYSIS OF THE BERT TEXT CLASSIFICATION ALGORITHM,” vol. 11, no. 3, pp. 537–546, 2024, doi: 10.25126/jtiik938093.
- [43] A. F. Abdillah, D. Purwitasari, S. Junita, and M. H. Purnomo, “Pengenalannya Entitas Biomedis Dalam Teks Konsultasi Kesehatan Biomedial Entity Recognition in Indonesian Online Health,” vol. 10, no. 1, pp. 131–140, 2023, doi: 10.25126/jtiik.2023106337.