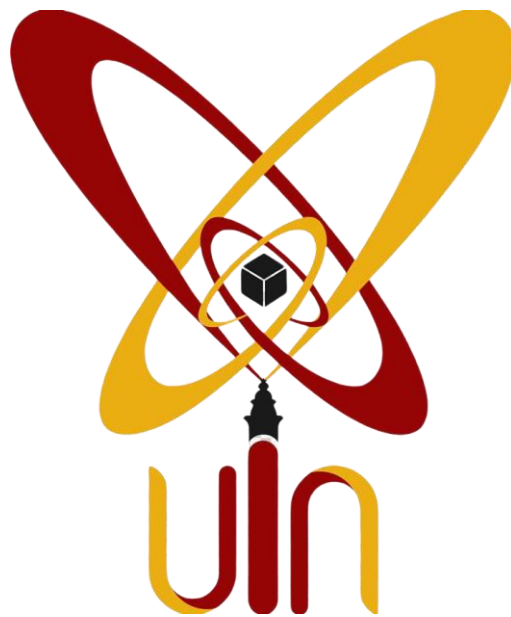


**ANALISIS KOMPARATIF SCIBERT, BIOBERT, DAN
PUBMEDBERT UNTUK NAMED ENTITY RECOGNATION
PADA LITERATUR RUMPUT LAUT**

Disusun untuk memenuhi salah satu syarat pelaksanaan Magang
pada Program Studi Informatika



Disusun oleh:

Nama : Nabila Nurasyifa

NIM : 231730020

**PROGRAM STUDI INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI
SULTAN MAULANA HASANUDDIN BANTEN
2026**

HALAMAN PENGESAHAN

Laporan hasil pelaksanaan PKL/Magang ini telah disusun oleh:

Nama : Nabila Nurasyifa
NIM : 231730020
Program Studi : Informatika
Judul Magang : Analisis Komparatif SciBERT, BioBERT, dan PubMedBERT untuk Named Entity Recognition pada Literatur Rumput Laut
Tempat Magang : Badan Riset dan Inovasi Nasional (BRIN)
Waktu Pelaksanaan : 18 Februari 2026 - 18 Juli 2026

Laporan ini telah diperiksa dan disahkan sebagai laporan hasil pelaksanaan Magang.

Bandung, 24 Juni 2026

Dosen Pembimbing Magang,



Ahmad Tabrani, M.T.I
NIP. 198509172018011002

Pembimbing Lapangan,



Dr. Dipl. Ing. (FH) Anne Parlina M.P.
NIP 197602141994032001

Mengetahui,

Ketua Program Studi Informatika



Ahmad Tabrani, M.T.I
NIP. 198509172018011002

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah SWT karena atas rahmat dan karunia-Nya, penulis dapat menyelesaikan laporan hasil pelaksanaan Magang yang berjudul “Analisis Komparatif SciBERT, BioBERT, dan PubMedBERT untuk *Named Entity Recognition* pada Literatur Rumput Laut” dengan baik.

Laporan ini disusun sebagai salah satu bentuk pertanggungjawaban atas kegiatan Magang yang telah dilaksanakan di Badan Riset dan Inovasi Nasional (BRIN) Melalui kegiatan ini, penulis memperoleh pengalaman, pengetahuan, serta pemahaman secara langsung mengenai dunia kerja, khususnya dalam bidang *kecerdasan buatan (Artificial Intelligence)*, *pengolahan bahasa alami (Natural Language Processing)*, dan *data science*.

Penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan bantuan, bimbingan, dan dukungan selama pelaksanaan Magang hingga penyusunan laporan ini, terutama kepada:

1. Bapak Ahmad Tabrani, M.T.I
2. Ibu Dr. Dipl. Ing. (FH) Anne Parlina M.P.
4. Badan Riset Inovasi Nasional (BRIN)
5. Semua pihak yang telah membantu dalam pelaksanaan Magang.

Penulis menyadari bahwa laporan ini masih memiliki kekurangan. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan agar laporan ini dapat menjadi lebih baik. Semoga laporan ini dapat memberikan manfaat bagi pembaca dan pihak-pihak yang membutuhkan.

Bandung, 24 Juni 2026



Nabila Nurasyifa

DAFTAR ISI

HALAMAN PENGESAHAN.....	i
KATA PENGANTAR	ii
DAFTAR ISI.....	iii
RINGKASAN	v
BAB I PENDAHULUAN	1
1.1 Latar Belakang Masalah.....	1
1.2 Rumusan Masalah	3
1.3 Tujuan Magang	3
1.3.1 Tujuan umum	3
1.4 Manfaat Magang	4
1.4.1 Manfaat teoritis	4
BAB II KAJIAN PUSTAKA	6
2.1 Kajian Pustaka.....	6
2.1.1 Pengertian NER.....	6
2.1.2 Konsep SciBERT, BioBERT, PubMedBERT	7
2.2 Kerangka Pikir Teoritis	8
BAB III METODE	10
3.1 Lokasi Magang.....	10
3.2 Desain Pelaksanaan Magang.....	10
3.3 Penentuan Objek Magang	11
3.4 Metode Pengumpulan Data	11
3.5 Instrumen Magang	11
3.6 Teknik Analisis Data.....	11

BAB IV HASIL DAN PEMBAHASAN	13
4.1 Deskripsi Kegiatan Magang	13
4.2 Permasalahan yang Ditemukan	13
4.3 Metode Penyelesaian Masalah	14
4.4 Hasil Pelaksanaan Magang	14
4.5 Pembahasan.....	19
4.6 Pelajaran yang Diperoleh	19
BAB V KESIMPULAN DAN SARAN.....	21
5.1 Kesimpulan	21
5.2 Keterbatasan.....	21
5.3 Implikasi.....	21
5.4 Saran.....	22
DAFTAR PUSTAKA	23
LAMPIRAN.....	24

RINGKASAN

Perkembangan penelitian mengenai rumput laut (*seaweed*) menghasilkan semakin banyak publikasi ilmiah yang memuat informasi penting, seperti nama spesies, lokasi penelitian, senyawa bioaktif, dan manfaatnya. Namun, sebagian besar informasi tersebut masih tersimpan dalam bentuk teks tidak terstruktur sehingga proses pencarian dan ekstraksi informasi secara manual menjadi kurang efisien. Kondisi tersebut menunjukkan perlunya metode yang mampu mengubah informasi pada literatur ilmiah menjadi data terstruktur. Salah satu pendekatan yang dapat digunakan adalah *Named Entity Recognition* (NER), yaitu teknik *Natural Language Processing* (NLP) untuk mengidentifikasi dan mengklasifikasikan entitas tertentu secara otomatis.

Kegiatan magang dilaksanakan di Pusat Riset Sains Data dan Informasi (PRSDI), Badan Riset dan Inovasi Nasional (BRIN). Objek kegiatan berupa ekstraksi informasi dari abstrak literatur ilmiah rumput laut yang diperoleh dari basis data Scopus. Data dikumpulkan melalui proses pengambilan abstrak artikel ilmiah, kemudian dianotasi menggunakan Label Studio dengan kategori entitas *Seaweed*, *Location*, *Compound*, dan *Benefit*. Hasil anotasi disimpan dalam format JSON, dilanjutkan dengan tahap *preprocessing*, kemudian digunakan untuk pelatihan dan evaluasi model SciBERT, BioBERT, dan PubMedBERT. Analisis data dilakukan secara deskriptif dengan membandingkan performa ketiga model menggunakan metrik *precision*, *recall*, *F1-score*, dan *accuracy*.

Hasil pelaksanaan magang menunjukkan bahwa ketiga model mampu melakukan tugas *Named Entity Recognition* (NER) dengan baik pada literatur ilmiah rumput laut. Berdasarkan hasil evaluasi, PubMedBERT memberikan performa terbaik dengan nilai *precision* sebesar 92,74%, *recall* sebesar 85,93%, *F1-score* sebesar 88,30%, dan *accuracy* sebesar 85,93%. Selain menghasilkan perbandingan performa model *Transformer* pada domain rumput laut, kegiatan magang juga memberikan pengalaman dalam proses pengumpulan data, anotasi dataset, pengembangan model, serta evaluasi hasil penelitian di bidang *Natural Language Processing*.

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Pelaksanaan magang merupakan salah satu kegiatan akademik yang bertujuan untuk memberikan pengalaman langsung kepada mahasiswa dalam dunia kerja. Melalui kegiatan ini, mahasiswa dapat menerapkan ilmu yang telah diperoleh selama perkuliahan sekaligus mengembangkan kompetensi sesuai dengan bidang keahlian pada lingkungan kerja profesional. Selain memberikan pengalaman praktis, kegiatan magang juga menjadi sarana untuk memahami proses kerja, meningkatkan kemampuan berpikir kritis, serta memperoleh pengalaman dalam menyelesaikan berbagai permasalahan yang dihadapi oleh instansi tempat pelaksanaan magang.

Kegiatan magang dilaksanakan di Pusat Riset Sains Data dan Informasi (PRSDI), Badan Riset dan Inovasi Nasional (BRIN), yang merupakan salah satu pusat riset nasional yang berfokus pada pengembangan ilmu pengetahuan dan teknologi, khususnya di bidang sains data, kecerdasan buatan (*Artificial Intelligence*), dan pengolahan bahasa alami (*Natural Language Processing*). Selama pelaksanaan magang, kegiatan yang dilakukan berfokus pada penelitian mengenai pengembangan metode ekstraksi informasi dari literatur ilmiah menggunakan teknik *Named Entity Recognition* (NER).

Perkembangan penelitian mengenai rumput laut (*seaweed*) mengalami peningkatan yang sangat pesat dalam beberapa tahun terakhir. Rumput laut menjadi salah satu sumber daya hayati yang memiliki potensi besar pada berbagai bidang, seperti pangan, kesehatan, farmasi, kosmetik, hingga lingkungan. Seiring meningkatnya aktivitas penelitian tersebut, jumlah publikasi ilmiah yang membahas berbagai aspek rumput laut juga terus bertambah. Informasi penting yang terkandung di dalam publikasi tersebut meliputi nama spesies rumput laut, lokasi penelitian, kandungan senyawa bioaktif, serta berbagai manfaat yang dimiliki oleh rumput laut. Namun, sebagian besar informasi tersebut masih tersimpan dalam bentuk teks tidak terstruktur sehingga proses pencarian,

identifikasi, maupun analisis informasi secara manual menjadi kurang efisien, terutama ketika jumlah literatur yang dianalisis sangat besar.

Kondisi tersebut menunjukkan perlunya suatu metode yang mampu melakukan ekstraksi informasi secara otomatis sehingga informasi yang terdapat pada literatur ilmiah dapat diubah menjadi data terstruktur yang lebih mudah dianalisis. Salah satu teknik yang banyak digunakan dalam bidang *Natural Language Processing* (NLP) untuk menyelesaikan permasalahan tersebut adalah *Named Entity Recognition* (NER). NER merupakan teknik yang digunakan untuk mengidentifikasi dan mengklasifikasikan entitas tertentu di dalam suatu teks ke dalam kategori yang telah ditentukan. Entitas yang digunakan dalam penelitian ini meliputi *Seaweed*, *Location*, *Compound*, dan *Benefit*, sehingga informasi penting mengenai rumput laut dapat diekstraksi secara otomatis dari literatur ilmiah.

Perkembangan model berbasis *Transformer* telah memberikan peningkatan performa yang signifikan pada berbagai tugas NLP, termasuk NER. Model seperti SciBERT, BioBERT, dan PubMedBERT dikembangkan melalui proses *pre-training* pada korpus ilmiah maupun biomedis sehingga memiliki kemampuan yang lebih baik dalam memahami istilah-istilah ilmiah dibandingkan model BERT umum. Meskipun demikian, sebagian besar penelitian sebelumnya masih berfokus pada dataset NER di bidang biomedis secara umum. Penelitian yang secara khusus membandingkan performa ketiga model tersebut pada literatur ilmiah rumput laut masih sangat terbatas. Kondisi ini menunjukkan adanya kebutuhan untuk mengidentifikasi model yang paling efektif dalam melakukan ekstraksi entitas pada domain rumput laut.

Berdasarkan kondisi tersebut, kegiatan penelitian selama pelaksanaan magang difokuskan pada pembangunan dataset *Named Entity Recognition* dari abstrak ilmiah rumput laut yang diperoleh dari basis data Scopus. Tahapan penelitian meliputi proses anotasi data menggunakan Label Studio, penerapan skema pelabelan BIO (*Beginning-Inside-Outside*), implementasi serta *fine-tuning* model SciBERT, BioBERT, dan PubMedBERT, hingga evaluasi performa masing-masing model menggunakan metrik *precision*, *recall*, *F1-score*, dan *accuracy*. Hasil penelitian diharapkan dapat memberikan informasi mengenai

model *Transformer* yang paling efektif dalam mengenali entitas pada literatur ilmiah rumput laut sekaligus mendukung pengembangan metode ekstraksi informasi pada domain tersebut.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam pelaksanaan magang adalah sebagai berikut.

1. Bagaimana pelaksanaan kegiatan magang di Pusat Riset Sains Data dan Informasi (PRSDI), Badan Riset dan Inovasi Nasional (BRIN)?
2. Bagaimana proses pengembangan metode *Named Entity Recognition* (NER) untuk ekstraksi informasi pada literatur ilmiah rumput laut selama pelaksanaan magang?
3. Bagaimana hasil implementasi dan perbandingan model SciBERT, BioBERT, dan PubMedBERT dalam mengenali entitas pada literatur ilmiah rumput laut?

1.3 Tujuan Magang

1.3.1 Tujuan umum

Tujuan umum pelaksanaan magang adalah memperoleh pengalaman kerja secara langsung di lingkungan penelitian serta mengembangkan pengetahuan, keterampilan, dan kompetensi di bidang sains data, kecerdasan buatan, dan *Natural Language Processing* melalui kegiatan penelitian di Pusat Riset Sains Data dan Informasi (PRSDI), Badan Riset dan Inovasi Nasional (BRIN).

1.3.2 Tujuan khusus

1. Mengetahui pelaksanaan kegiatan penelitian di Pusat Riset Sains Data dan Informasi (PRSDI), Badan Riset dan Inovasi Nasional (BRIN).
2. Mempelajari tahapan pengembangan *Named Entity Recognition* (NER) untuk ekstraksi informasi pada literatur ilmiah rumput laut,

mulai dari proses anotasi data, *preprocessing*, pelatihan model, hingga evaluasi.

3. Mengimplementasikan model SciBERT, BioBERT, dan PubMedBERT pada tugas *Named Entity Recognition* untuk mengenali entitas *Seaweed*, *Location*, *Compound*, dan *Benefit*.
4. Menganalisis dan membandingkan performa model SciBERT, BioBERT, dan PubMedBERT berdasarkan metrik *precision*, *recall*, *F1-score*, dan *accuracy*..

1.4 Manfaat Magang

1.4.1 Manfaat teoritis

Pelaksanaan magang diharapkan dapat memberikan kontribusi terhadap pengembangan ilmu pengetahuan, khususnya pada bidang *Natural Language Processing* (NLP) dan *Named Entity Recognition* (NER). Selain itu, hasil penelitian diharapkan dapat menjadi referensi bagi penelitian selanjutnya yang berkaitan dengan ekstraksi informasi pada literatur ilmiah, terutama dalam penerapan model *Transformer* seperti SciBERT, BioBERT, dan PubMedBERT pada domain rumput laut.

1.4.2 Manfaat praktis

1. Bagi mahasiswa, magang memberikan pengalaman dalam penerapan ilmu yang telah diperoleh selama perkuliahan melalui keterlibatan pada kegiatan penelitian di bidang sains data, kecerdasan buatan, dan *Natural Language Processing*. Selain itu, magang juga meningkatkan pemahaman mengenai proses penelitian, mulai dari anotasi data, pengolahan dataset, pelatihan model, hingga evaluasi hasil penelitian.
2. Bagi Badan Riset dan Inovasi Nasional (BRIN), hasil penelitian diharapkan dapat menjadi salah satu referensi dalam pengembangan metode ekstraksi informasi dari literatur ilmiah rumput laut serta memberikan gambaran mengenai performa model SciBERT, BioBERT, dan PubMedBERT pada tugas *Named Entity Recognition*.

3. Bagi Universitas, hasil magang diharapkan dapat memperkuat kerja sama antara perguruan tinggi dengan instansi mitra, sekaligus menjadi referensi bagi pelaksanaan kegiatan magang dan penelitian selanjutnya, khususnya pada bidang *Artificial Intelligence*, *Natural Language Processing*, dan *text mining*.

BAB II

KAJIAN PUSTAKA

2.1 Kajian Pustaka

Kajian pustaka merupakan landasan teoritis yang digunakan untuk mendukung pelaksanaan penelitian selama kegiatan magang. Landasan teori berfungsi sebagai dasar dalam memahami konsep, metode, dan model yang digunakan sehingga proses analisis serta pembahasan hasil penelitian dapat dilakukan secara sistematis. Pada laporan ini, kajian pustaka difokuskan pada teori mengenai *Named Entity Recognition* (NER) sebagai metode ekstraksi informasi dari teks serta model SciBERT, BioBERT, dan PubMedBERT yang digunakan untuk mengenali entitas pada literatur ilmiah rumput laut. Teori-teori tersebut dipilih karena memiliki keterkaitan langsung dengan tahapan penelitian, mulai dari proses anotasi data hingga evaluasi performa model.

2.1.1 Pengertian NER

Named Entity Recognition (NER) merupakan salah satu tugas dalam *Natural Language Processing* (NLP) yang bertujuan untuk mengidentifikasi dan mengklasifikasikan entitas tertentu ke dalam kategori yang telah ditentukan [1]. Teknik ini berperan penting dalam proses ekstraksi informasi karena mampu mengubah data yang masih berbentuk teks tidak terstruktur menjadi informasi yang lebih terorganisasi sehingga lebih mudah diproses dan dianalisis.

Penerapan *NER* telah berkembang pesat pada berbagai bidang, khususnya domain ilmiah dan biomedis yang memiliki jumlah publikasi sangat besar. Pemanfaatan *NER* memungkinkan proses identifikasi informasi penting dilakukan secara otomatis sehingga dapat meningkatkan efisiensi pengolahan dokumen ilmiah.[2]

Pada penelitian ini, proses *Named Entity Recognition* digunakan untuk mengenali empat kategori entitas, yaitu *Seaweed*, *Location*, *Compound*, dan *Benefit*. Keempat kategori tersebut dipilih karena merepresentasikan informasi utama yang terkandung pada abstrak ilmiah rumput laut sehingga

dapat mendukung proses ekstraksi informasi secara otomatis. Proses anotasi dilakukan menggunakan skema BIO (*Beginning-Inside-Outside*), yaitu metode pelabelan yang membedakan token awal suatu entitas (*Beginning*), token lanjutan (*Inside*), serta token yang tidak termasuk ke dalam entitas (*Outside*). Penggunaan skema BIO membantu model dalam mengenali batas setiap entitas sehingga proses identifikasi menjadi lebih akurat.

2.1.2 Konsep SciBERT, BioBERT, PubMedBERT

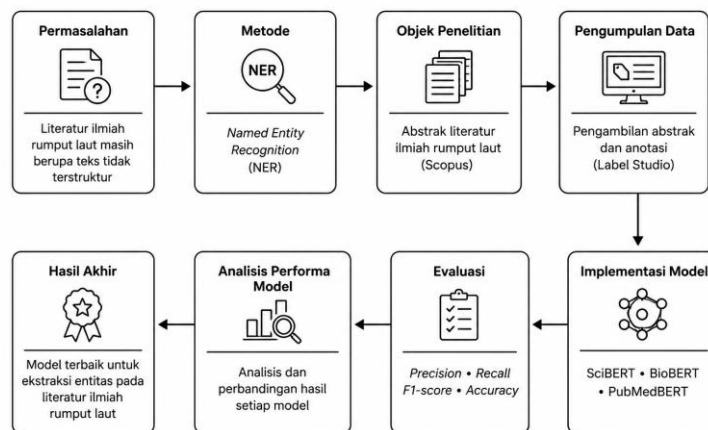
SciBERT, BioBERT, dan PubMedBERT merupakan model bahasa berbasis arsitektur *Transformer* yang dikembangkan untuk meningkatkan performa berbagai tugas *Natural Language Processing*, termasuk *Named Entity Recognition*. Arsitektur *Transformer* memanfaatkan mekanisme *self-attention* sehingga mampu memahami hubungan antarkata dalam suatu kalimat secara lebih efektif dibandingkan pendekatan sebelumnya [3]. Pengembangan BERT (*Bidirectional Encoder Representations from Transformers*) selanjutnya memungkinkan model menghasilkan representasi bahasa yang lebih kontekstual melalui proses *bidirectional pre-training* [4].

SciBERT dikembangkan menggunakan korpus publikasi ilmiah dari berbagai disiplin ilmu sehingga mampu memahami terminologi ilmiah dengan lebih baik dibandingkan BERT umum [5]. Sementara itu, BioBERT merupakan pengembangan BERT yang dilatih kembali menggunakan korpus biomedis, seperti PubMed dan *PubMed Central*, sehingga memberikan peningkatan performa pada berbagai tugas *biomedical text mining*, termasuk *Named Entity Recognition* [6].

Berbeda dengan BioBERT, PubMedBERT dikembangkan melalui proses *pre-training* sejak awal menggunakan korpus PubMed tanpa memanfaatkan bobot awal dari BERT umum. Pendekatan tersebut menghasilkan representasi bahasa yang lebih spesifik terhadap domain biomedis sehingga mampu meningkatkan performa pada berbagai tugas pemrosesan bahasa alami di bidang kesehatan dan ilmu hayati [7].

Pada penelitian ini, ketiga model dibandingkan untuk mengetahui pengaruh perbedaan korpus *pre-training* terhadap kemampuan model dalam mengenali entitas *Seaweed*, *Location*, *Compound*, dan *Benefit* pada literatur ilmiah rumput laut.

2.2 Kerangka Pikir Teoritis



Gambar 1. Kerangka Pikir Teoritis

Kerangka pikir teoritis menggambarkan hubungan antara permasalahan, landasan teori, objek penelitian, metode yang digunakan, serta hasil yang diharapkan selama pelaksanaan magang. Permasalahan utama dalam penelitian ini adalah masih banyaknya informasi penting pada literatur ilmiah rumput laut yang tersimpan dalam bentuk teks tidak terstruktur sehingga proses pencarian, identifikasi, dan analisis informasi secara manual memerlukan waktu yang relatif lama. Kondisi tersebut menunjukkan perlunya suatu metode yang mampu melakukan ekstraksi informasi secara otomatis agar informasi yang terdapat pada literatur ilmiah dapat diubah menjadi data yang lebih terstruktur.

Pendekatan yang digunakan untuk menyelesaikan permasalahan tersebut adalah *Named Entity Recognition (NER)*, yaitu salah satu tugas dalam *Natural Language Processing* yang bertujuan mengidentifikasi dan mengklasifikasikan entitas pada suatu teks. Objek penelitian berupa abstrak literatur ilmiah rumput laut yang diperoleh dari basis data Scopus. Data dikumpulkan melalui proses

pengambilan abstrak ilmiah, kemudian dilakukan anotasi menggunakan Label Studio dengan menerapkan skema BIO (*Beginning-Inside-Outside*) sehingga menghasilkan dataset berlabel yang siap digunakan pada proses pelatihan model.

Tahap berikutnya adalah implementasi tiga model berbasis *Transformer*, yaitu SciBERT, BioBERT, dan PubMedBERT. Setiap model dilatih menggunakan dataset yang sama sehingga perbedaan performa yang dihasilkan dipengaruhi oleh karakteristik masing-masing model. Selanjutnya, hasil prediksi dievaluasi menggunakan metrik *precision*, *recall*, *F1-score*, dan *accuracy* untuk mengetahui kemampuan setiap model dalam mengenali entitas *Seaweed*, *Location*, *Compound*, dan *Benefit*. Hasil evaluasi tersebut kemudian dianalisis untuk menentukan model yang memiliki performa terbaik dalam melakukan ekstraksi informasi pada literatur ilmiah rumput laut.

Kerangka pikir teoritis tersebut menggambarkan bahwa teori *Named Entity Recognition* menjadi dasar dalam proses ekstraksi informasi, sedangkan SciBERT, BioBERT, dan PubMedBERT berperan sebagai model yang dibandingkan untuk memperoleh performa terbaik.

BAB III

METODE

3.1 Lokasi Magang

Kegiatan Magang dilaksanakan di Badan Riset dan Inovasi Nasional (BRIN), tepatnya pada Organisasi Riset Elektronika dan Informatika (OREI), Pusat Riset Sains Data dan Informasi (PRSDI). Pusat riset ini memiliki bidang penelitian yang berfokus pada sains data, kecerdasan buatan (*Artificial Intelligence*), pembelajaran mesin (*Machine Learning*), pemrosesan bahasa alami (*Natural Language Processing*), serta temu kembali informasi (*Information Retrieval*).

Alamat instansi berada di Jl. Tamansari No. 71, Lebak Siliwangi, Kecamatan Coblong, Kota Bandung, Jawa Barat 40132, Indonesia. Pelaksanaan Magang dilakukan pada Kelompok Riset Temu Kembali Informasi (*Information Retrieval Research Group*) di bawah Pusat Riset Sains Data dan Informasi (PRSDI) kemudian dilaksanakan selama kurang lebih lima bulan, yaitu mulai 18 Februari 2026 hingga 31 Juli 2026.

3.2 Desain Pelaksanaan Magang

Pelaksanaan magang menggunakan pendekatan deskriptif melalui keterlibatan dalam kegiatan penelitian di bidang *Natural Language Processing* (NLP). Rangkaian kegiatan meliputi pengumpulan data, anotasi dataset menggunakan Label Studio, *preprocessing*, implementasi model SciBERT, BioBERT, dan PubMedBERT, serta evaluasi performa model menggunakan metrik *precision*, *recall*, *F1-score*, dan *accuracy*. Pendekatan ini dipilih karena sesuai dengan tujuan penelitian, yaitu membandingkan performa ketiga model pada tugas *Named Entity Recognition* (NER).

3.3 Penentuan Objek Magang

Objek kegiatan magang adalah proses ekstraksi informasi dari literatur ilmiah rumput laut menggunakan metode *Named Entity Recognition* (NER). Data yang digunakan berupa abstrak artikel ilmiah yang diperoleh dari Scopus dan dianotasi ke dalam empat kategori entitas, yaitu *Seaweed*, *Location*, *Compound*, dan *Benefit*. Selanjutnya, dataset digunakan untuk implementasi serta evaluasi model SciBERT, BioBERT, dan PubMedBERT..

3.4 Metode Pengumpulan Data

Metode pengumpulan data dilakukan melalui pengumpulan abstrak artikel ilmiah mengenai rumput laut dari basis data Scopus menggunakan kata kunci yang telah ditentukan. Abstrak yang diperoleh kemudian diseleksi untuk memastikan kesesuaiannya dengan topik penelitian. Selanjutnya, setiap abstrak dianotasi menggunakan Label Studio dengan empat kategori entitas, yaitu *Seaweed*, *Location*, *Compound*, dan *Benefit*. Hasil anotasi disimpan dalam format JSON sebagai dataset awal, kemudian dilakukan proses *preprocessing* untuk menyesuaikan format data sebelum digunakan sebagai data pelatihan dan evaluasi model SciBERT, BioBERT, dan PubMedBERT.

3.5 Instrumen Magang

Instrumen yang digunakan selama kegiatan meliputi laptop sebagai perangkat utama, buku catatan untuk mencatat hasil diskusi, serta beberapa perangkat lunak pendukung, yaitu *Python* sebagai bahasa pemrograman, *Google Colab* sebagai lingkungan pengembangan, *Kaggle* sebagai sumber *dataset*, *GitHub* sebagai media referensi implementasi, dan *Label Studio* sebagai alat anotasi data dalam proses pengembangan model *Named Entity Recognition* (NER).

3.6 Teknik Analisis Data

Teknik analisis data dilakukan secara deskriptif dengan menganalisis hasil implementasi model *Named Entity Recognition* (NER) pada dataset literatur ilmiah rumput laut. Dataset yang telah melalui proses anotasi dan *preprocessing*

digunakan untuk pelatihan dan pengujian model SciBERT, BioBERT, dan PubMedBERT. Selanjutnya, hasil prediksi dari setiap model dievaluasi menggunakan metrik *precision*, *recall*, *F1-score*, dan *accuracy* untuk mengukur kemampuan model dalam mengenali entitas *Seaweed*, *Location*, *Compound*, dan *Benefit*. Hasil evaluasi kemudian dibandingkan dan diinterpretasikan untuk mengetahui model yang memiliki performa terbaik pada tugas ekstraksi entitas dalam literatur ilmiah rumput laut.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Deskripsi Kegiatan Magang

Tabel 1. Kegiatan Magang

No	Tanggal	Kegiatan	Keterangan
1	Februari – Maret 2026	<i>Research & System Design</i>	Melakukan studi literatur terkait <i>Named Entity Recognition</i> (NER), menentukan skema entitas, serta menyusun rancangan sistem yang akan dikembangkan.
2	Maret – April 2026	<i>Dataset & Annotation</i>	Melakukan pembersihan (<i>cleaning</i>) dataset dan proses anotasi data menggunakan Label Studio sesuai skema entitas yang telah ditentukan.
3	April – Juni 2026	<i>NER Model Development</i>	Melakukan pelatihan (<i>training</i>), pengujian, dan evaluasi model NER menggunakan pendekatan <i>Transformer-based</i> .
4	Juni – Juli 2026	Dokumentasi	Menyusun laporan PKL, menyelesaikan penulisan <i>paper</i> penelitian (<i>paper finalization</i>), serta mendokumentasikan perangkat lunak yang dikembangkan.

Sumber: Data diolah oleh penulis, 2026

4.2 Permasalahan yang Ditemukan

Selama pelaksanaan magang, ditemukan beberapa permasalahan dalam proses pengembangan *Named Entity Recognition* (NER) pada literatur ilmiah rumput laut. Permasalahan pertama adalah data yang diperoleh dari abstrak artikel ilmiah masih berbentuk teks tidak terstruktur sehingga informasi penting, seperti nama rumput laut, lokasi penelitian, senyawa, dan manfaat, belum dapat dimanfaatkan secara langsung.

Permasalahan berikutnya adalah belum tersedianya dataset *Named Entity Recognition* khusus pada domain rumput laut. Kondisi tersebut mengharuskan proses anotasi dilakukan secara manual menggunakan Label Studio sehingga

memerlukan ketelitian untuk menghasilkan dataset yang sesuai dengan kebutuhan penelitian.

Selain itu, pemilihan model yang paling sesuai untuk mengenali entitas pada literatur ilmiah rumput laut juga menjadi tantangan. Meskipun SciBERT, BioBERT, dan PubMedBERT sama-sama dikembangkan untuk domain ilmiah dan biomedis, performa masing-masing model terhadap dataset rumput laut belum diketahui sehingga diperlukan proses implementasi dan evaluasi untuk memperoleh model dengan hasil terbaik.

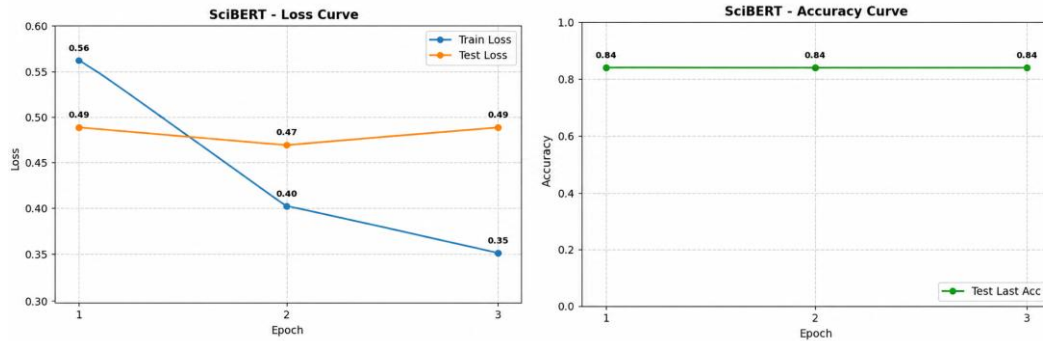
4.3 Metode Penyelesaian Masalah

Penyelesaian permasalahan dilakukan melalui beberapa tahapan. Abstrak artikel ilmiah dikumpulkan dari basis data Scopus, kemudian dilakukan proses anotasi menggunakan Label Studio dengan kategori entitas *Seaweed*, *Location*, *Compound*, dan *Benefit*. Hasil anotasi diekspor dalam format JSON dan diproses lebih lanjut melalui tahap *preprocessing* sebelum digunakan sebagai dataset pelatihan dan pengujian.

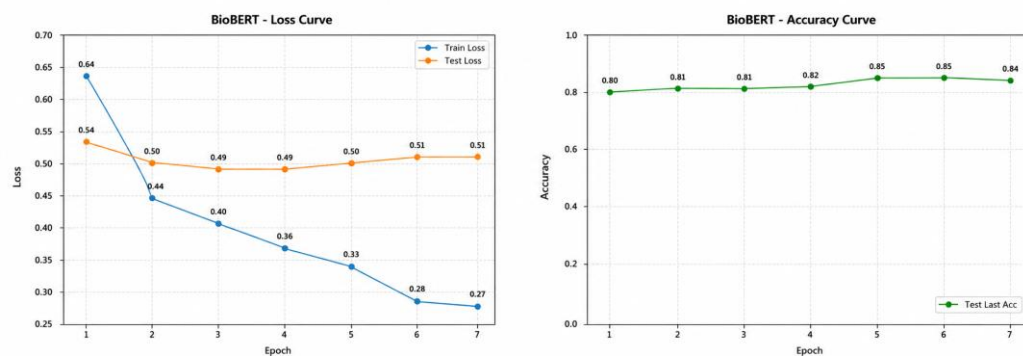
Selanjutnya, dilakukan implementasi model SciBERT, BioBERT, dan PubMedBERT pada dataset yang sama. Performa setiap model dievaluasi menggunakan metrik *precision*, *recall*, *F1-score*, dan *accuracy*, kemudian hasilnya dibandingkan untuk mengidentifikasi model yang paling efektif dalam mengenali entitas pada literatur ilmiah rumput laut.

4.4 Hasil Pelaksanaan Magang

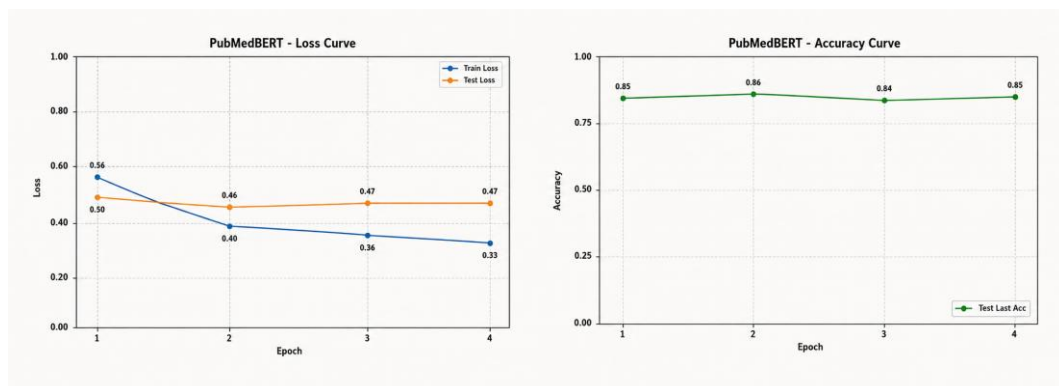
Selama pelaksanaan magang telah dilakukan implementasi tiga model berbasis *Transformer*, yaitu SciBERT, BioBERT, dan PubMedBERT, pada tugas *Named Entity Recognition* (NER) untuk ekstraksi informasi dari literatur ilmiah rumput laut. Proses implementasi meliputi tahap pelatihan (*training*) dan evaluasi model menggunakan dataset yang telah dianotasi. Hasil pelaksanaan magang disajikan dalam bentuk grafik *training loss*, grafik *validation loss*, grafik *accuracy*, *confusion matrix*, serta tabel hasil evaluasi yang mencakup nilai *precision*, *recall*, *F1-score*, dan *accuracy* sebagai dasar untuk membandingkan performa masing-masing model.



Gambar 1. Visualisasi loss curve & accuracy curve SciBERT



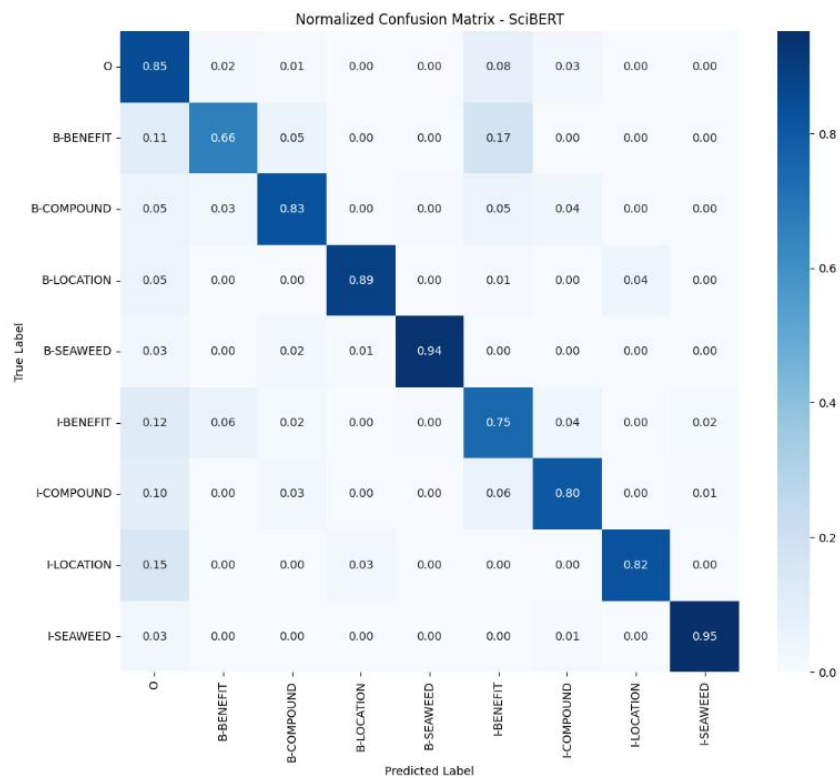
Gambar 2. Visualisasi loss curve & accuracy curve BioBERT



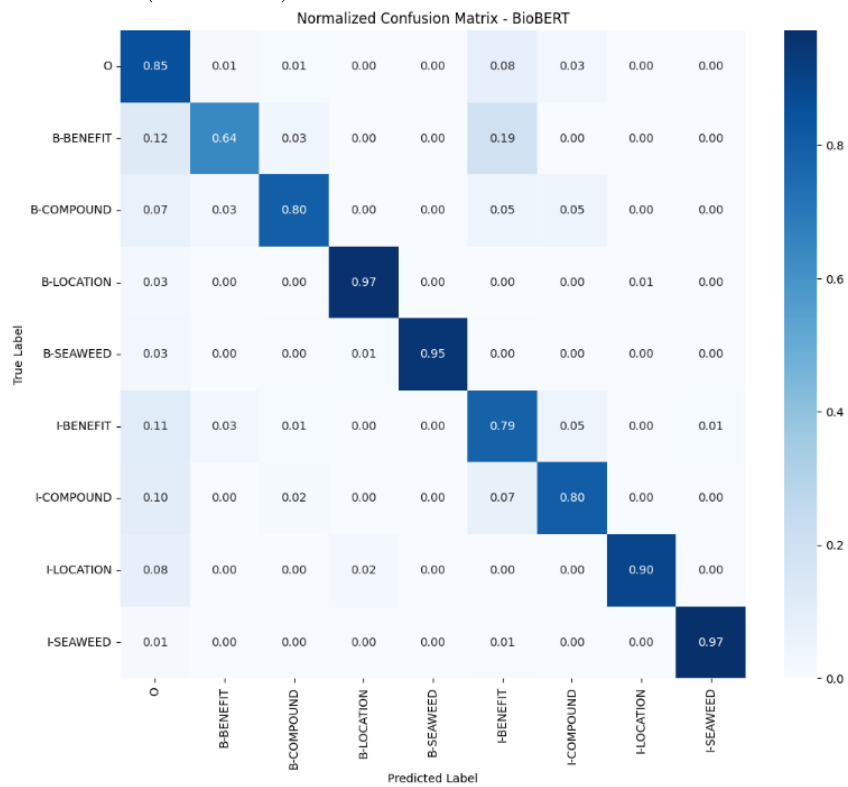
Gambar 3. Visualisasi loss curve & accuracy curve PubMedBERT

Berdasarkan Gambar 1, Gambar .2, dan Gambar .3, ketiga model menunjukkan proses pelatihan yang berjalan dengan baik. Hal ini ditunjukkan oleh nilai *training loss* yang terus menurun pada setiap *epoch*, sedangkan *validation loss* relatif stabil selama proses *fine-tuning*. Selain itu, nilai *accuracy* mengalami peningkatan hingga mencapai performa terbaik pada masing-masing

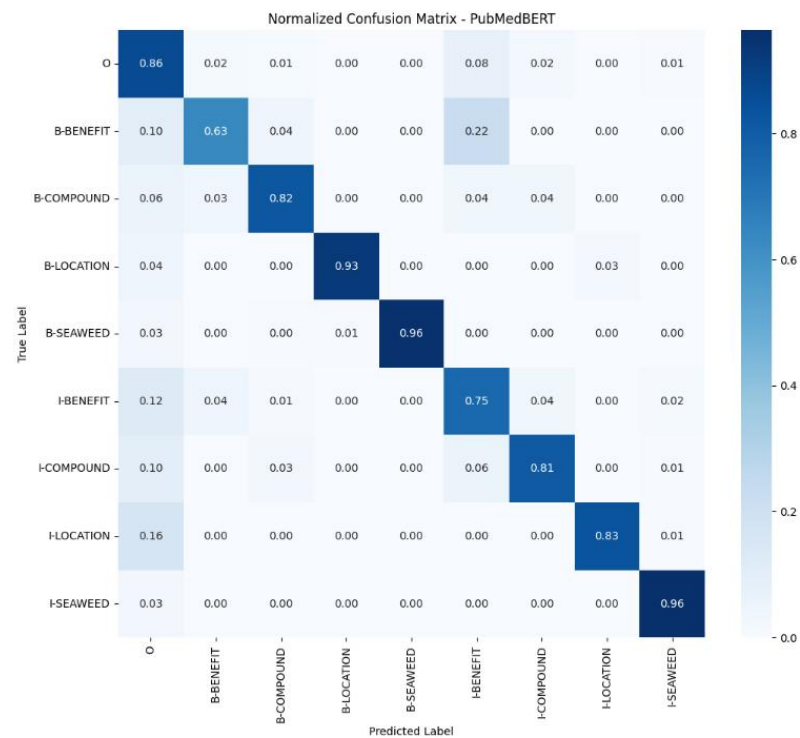
model. SciBERT mencapai performa terbaik pada *epoch* pertama, BioBERT pada *epoch* keenam, sedangkan PubMedBERT pada *epoch* kedua. Hasil tersebut menunjukkan bahwa ketiga model mampu mempelajari pola pada dataset *Named Entity Recognition* (NER), meskipun memiliki karakteristik proses konvergensi yang berbeda.



Gambar 4. Confusion Matrix SciBERT



Gambar 5. Confusion Matrix BioBERT



Gambar 6. Confusion Matrix PuMedBERT

Gambar 4, Gambar.5, dan Gambar 6 menunjukkan *confusion matrix* dari model SciBERT, BioBERT, dan PubMedBERT. Secara umum, ketiga model memiliki kemampuan yang baik dalam mengenali entitas *Seaweed* dan *Location*, yang ditunjukkan oleh tingginya tingkat klasifikasi benar pada kedua kategori tersebut. Sebaliknya, entitas *Benefit* masih menjadi kategori yang paling sulit dikenali sehingga masih ditemukan beberapa kesalahan klasifikasi pada seluruh model. Meskipun demikian, hasil *confusion matrix* menunjukkan bahwa ketiga model telah mampu melakukan klasifikasi entitas dengan baik pada literatur ilmiah rumput laut.

Tabel 1. Perbandingan Kinerja Model NER

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
SciBERT	92.38	84.46	87.19	84.46
BioBERT	91.29	84.82	86.97	84.82
PubMedBERT	92.74	85.93	88.30	85.93

Berdasarkan Tabel 1, ketiga model menunjukkan performa yang baik pada tugas *Named Entity Recognition* (NER). PubMedBERT memperoleh hasil terbaik dengan nilai *precision* sebesar 92,74%, *recall* sebesar 85,93%, *F1-score* sebesar 88,30%, dan *accuracy* sebesar 85,93%. Hasil tersebut menunjukkan bahwa PubMedBERT lebih efektif dalam mengenali entitas pada literatur ilmiah rumput laut dibandingkan SciBERT dan BioBERT.

Tabel 2. Perbandingan F1-Score Setiap Kelas Entitas

Entity Class	SciBERT (%)	BioBERT (%)	PubMedBERT (%)
Seaweed	94.17	95.00	95.74
Location	85.37	90.80	92.22
Compound	80.40	81.87	79.30
Benefit	83.84	82.58	83.14

Berdasarkan Tabel 2, setiap model menunjukkan keunggulan pada kategori entitas yang berbeda. PubMedBERT memperoleh *F1-score* tertinggi pada entitas *Seaweed* dan *Location*, BioBERT memberikan performa terbaik pada entitas *Compound*, sedangkan SciBERT memperoleh *F1-score* tertinggi pada entitas

Benefit. Meskipun demikian, secara keseluruhan PubMedBERT tetap menunjukkan performa terbaik pada tugas *Named Entity Recognition* (NER) untuk literatur ilmiah rumput laut.

4.5 Pembahasan

Hasil eksperimen menunjukkan bahwa penggunaan model transformer berbasis domain memberikan pengaruh terhadap kinerja Named Entity Recognition (NER) pada literatur ilmiah rumput laut. PubMedBERT memperoleh performa terbaik dibandingkan SciBERT dan BioBERT, yang menunjukkan bahwa representasi bahasa yang diperoleh dari korpus biomedis mampu mendukung proses pengenalan entitas pada domain ilmiah.

Berdasarkan evaluasi setiap kelas entitas, kategori Benefit menjadi entitas yang paling menantang untuk dikenali oleh seluruh model. Hal ini disebabkan oleh variasi penyampaian informasi mengenai manfaat yang memiliki konteks lebih abstrak dan sering menyerupai kalimat umum. Sebaliknya, entitas Seaweed memperoleh nilai F1-score tertinggi pada seluruh model (>94%), yang menunjukkan bahwa istilah spesies dan terminologi rumput laut lebih mudah dipelajari oleh model karena memiliki pola penyebutan yang lebih konsisten..

Secara keseluruhan, penelitian ini menunjukkan bahwa SciBERT, BioBERT, dan PubMedBERT mampu diterapkan untuk ekstraksi entitas pada literatur rumput laut, dengan PubMedBERT sebagai model dengan kinerja terbaik.

4.6 Pelajaran yang Diperoleh

Pelaksanaan Magang yang dilaksanakan di BRIN memberikan pengetahuan dan pengalaman mengenai proses penelitian di bidang *Natural Language Processing* (NLP), khususnya pengembangan model *Named Entity Recognition* (NER). Selain itu, diperoleh keterampilan dalam pengolahan data, anotasi menggunakan *Label Studio*, pengembangan model menggunakan *Python* dan *Google Colab*, serta pemanfaatan *GitHub* sebagai media kolaborasi dan referensi implementasi.

Dari sisi sikap kerja, kegiatan Magang ini meningkatkan kedisiplinan, tanggung jawab, kemampuan berkomunikasi, serta kerja sama dalam tim riset. Pengalaman berdiskusi dengan pembimbing dan anggota tim juga memberikan wawasan mengenai penyelesaian masalah secara sistematis serta pentingnya kolaborasi dalam pelaksanaan penelitian.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil pelaksanaan magang di Pusat Riset Sains Data dan Informasi (PRSDI), Badan Riset dan Inovasi Nasional (BRIN), kegiatan magang telah memberikan pemahaman mengenai proses pengembangan *Named Entity Recognition* (NER) untuk ekstraksi informasi pada literatur ilmiah rumput laut. Rangkaian kegiatan yang dilakukan meliputi pengumpulan abstrak artikel ilmiah dari Scopus, anotasi dataset menggunakan Label Studio, *preprocessing* data, implementasi model SciBERT, BioBERT, dan PubMedBERT, serta evaluasi performa model menggunakan metrik *precision*, *recall*, *F1-score*, dan *accuracy*. Hasil evaluasi menunjukkan bahwa ketiga model mampu mengenali entitas pada literatur ilmiah rumput laut dengan baik, namun PubMedBERT memberikan performa terbaik dengan nilai *precision* sebesar 92,74%, *recall* sebesar 85,93%, *F1-score* sebesar 88,30%, dan *accuracy* sebesar 85,93%. Hasil tersebut menunjukkan bahwa PubMedBERT lebih efektif dalam melakukan ekstraksi entitas pada literatur ilmiah rumput laut dibandingkan SciBERT dan BioBERT.

5.2 Keterbatasan

Selama pelaksanaan Magang, terdapat beberapa keterbatasan, seperti waktu pelaksanaan yang terbatas, proses anotasi data yang membutuhkan ketelitian dan waktu yang cukup lama, serta perlunya beberapa kali pelatihan model untuk memperoleh performa yang optimal. Selain itu, kualitas dan jumlah data juga memengaruhi hasil pengembangan model.

5.3 Implikasi

Hasil Magang ini memberikan manfaat dalam penerapan teknologi *Natural Language Processing* (NLP), khususnya *Named Entity Recognition* (NER), untuk membantu proses ekstraksi informasi dari literatur ilmiah secara otomatis. Selain itu, kegiatan ini juga memberikan pengalaman praktis yang dapat mendukung

pengembangan kompetensi akademik dan profesional di bidang sains data dan kecerdasan buatan.

5.4 Saran

Bagi instansi, diharapkan dapat terus memberikan bimbingan dan kesempatan bagi mahasiswa untuk terlibat dalam kegiatan penelitian. Selain itu, penyampaian arahan, informasi, maupun masukan secara lebih terstruktur dan berkesinambungan diharapkan dapat semakin mendukung kelancaran proses Magang serta membantu mahasiswa dalam memahami setiap tahapan kegiatan dengan lebih optimal.

Bagi pihak kampus, diharapkan dapat terus memperkuat kerja sama dengan instansi penelitian serta meningkatkan koordinasi dan komunikasi terkait pelaksanaan Magang. Penyampaian informasi yang lebih terstruktur dan tepat waktu diharapkan dapat meminimalkan perubahan jadwal atau informasi yang bersifat mendadak, sehingga proses persiapan dan pelaksanaan Magang dapat berjalan lebih efektif dan optimal.

DAFTAR PUSTAKA

- [1] J. Li, A. Sun, J. Han and C. Li, A Survey on Deep Learning for Named Entity Recognition, *IEEE Trans. Knowl. Data Eng.* 34 (2022) 50–70.
- [2] D. Rebholz-Schuhmann, A. Oellrich and R. Hoehndorf, Text-mining solutions for biomedical research: enabling integrative biology, *Nature Reviews Genetics* 13 (2012) 829–839.
- [3] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proc. NAACL-HLT 2019 (Association for Computational Linguistics, Minneapolis, 2019)*, pp. 4171–4186.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention Is All You Need, in *Advances in Neural Information Processing Systems* 30 (2017), pp. 5998–6008.
- [5] I. Beltagy, K. Lo and A. Cohan, SciBERT: A Pretrained Language Model for Scientific Text, in *Proc. EMNLP-IJCNLP 2019, Hong Kong (Association for Computational Linguistics, 2019)*, pp. 3615–3620.
- [6] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining, *Bioinformatics* 36 (2020) 1234–1240.
- [7] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao and H. Poon, Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing, *ACM Trans. Comput. Healthc.* 3 (2021) 1–23.

LAMPIRAN



