BAB 2
LANDASAN TEORI

2.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) adalah cabang dari kecerdasan buatan yang mempelajari bagaimana komputer dapat memahami, menganalisis, dan menghasilkan bahasa manusia secara otomatis [6]. NLP memadukan linguistik, ilmu komputer, dan matematika statistik untuk membangun sistem yang mampu menafsirkan makna dari teks, suara, atau dokumen [6, 5, 1]. Di era digital saat ini, NLP banyak digunakan untuk aplikasi seperti pencarian informasi, penerjemahan mesin, analisis sentimen, dan ekstraksi informasi dari teks tidak terstruktur [4].

Dalam konteks kesehatan, NLP memiliki peran penting untuk mengolah data klinis yang sebagian besar masih berada dalam format teks bebas, seperti catatan dokter, hasil laboratorium, dan laporan medis [1]. Pengolahan data ini secara manual akan membutuhkan waktu, biaya, dan tenaga ahli yang tidak sedikit. Oleh sebab itu, NLP menjadi solusi yang menjanjikan untuk melakukan ekstraksi otomatis informasi klinis, mulai dari identifikasi diagnosis, deteksi gejala, hingga ekstraksi data demografis pasien.

Namun, penerapan NLP pada teks medis, khususnya di Indonesia, menghadapi sejumlah tantangan. Keragaman istilah medis, penggunaan bahasa campuran (Indonesia, Inggris, Latin), serta banyaknya singkatan membuat proses ekstraksi informasi menjadi tidak sederhana [8]. Selain itu, data medis cenderung memiliki struktur dan gaya penulisan yang berbeda-beda di setiap institusi kesehatan. Diperlukan model yang mampu memahami konteks bahasa secara lebih dalam agar dapat menghasilkan ekstraksi informasi yang akurat dan dapat diandalkan untuk mendukung pengambilan keputusan klinis [15, 4].

Berbagai studi telah menunjukkan bahwa NLP memberikan dampak positif dalam meningkatkan efisiensi manajemen data medis dan memperbaiki mutu layanan kesehatan. Teknologi NLP yang terus berkembang, khususnya dengan hadirnya model *deep learning* seperti BERT, membuka peluang lebih besar untuk menghasilkan sistem ekstraksi informasi medis yang lebih andal dan adaptif terhadap karakteristik data lokal [12, 13, 14].

2.2 Named Entity Recognition (NER)

Named Entity Recognition (NER) merupakan salah satu teknik utama dalam NLP yang bertujuan untuk mengidentifikasi dan mengklasifikasikan entitas penting dalam teks, seperti nama orang, lokasi, organisasi, serta istilah medis seperti diagnosis, obat, atau prosedur [25]. Dalam bidang kesehatan, NER menjadi komponen krusial karena mampu mengekstraksi informasi yang relevan dari catatan medis elektronik, sehingga mendukung pelaporan, penelitian, dan analisis epidemiologi [1, 26].

Metode NER tradisional di Indonesia umumnya masih berbasis *rule-based* dan *Conditional Random Fields* (CRF). Metode ini sangat tergantung pada kamus istilah (dictionary) atau aturan yang telah ditetapkan sebelumnya, sehingga performanya cenderung menurun ketika dihadapkan pada variasi penulisan diagnosis, penggunaan singkatan, dan perubahan konteks kalimat [8, 9]. Selain itu,

metode konvensional tidak mampu menangkap relasi antar kata dalam satu entitas diagnosis yang kompleks, seperti ”diabetes mellitus tipe dua” atau ”hipertensi esensial primer”.

Tantangan lain dalam penerapan NER pada data medis Indonesia adalah kekayaan istilah lokal dan penggunaan bahasa campuran dalam catatan medis. Penulisan diagnosis dapat menggunakan bahasa Indonesia, Latin, atau istilah Inggris, bahkan sering kali bercampur di satu kalimat. Hal ini menyebabkan hasil ekstraksi entitas menjadi tidak konsisten jika hanya mengandalkan metode berbasis aturan. Oleh sebab itu, dibutuhkan pendekatan yang mampu memahami konteks dan struktur bahasa secara mendalam.

Seiring dengan perkembangan teknologi *deep learning*, model NER berbasis *neural network*, khususnya arsitektur *transformer* seperti BERT, semakin banyak diadopsi. Model ini memiliki keunggulan dalam memahami konteks kata secara dinamis di seluruh bagian teks. Penelitian terbaru di Indonesia menunjukkan bahwa model berbasis BERT secara signifikan meningkatkan performa NER pada rekam medis elektronik, terutama dalam mengenali diagnosis yang kompleks dan beragam [15].

2.3 Multilingual Bert (mBERT)

Multilingual BERT (mBERT) adalah model *deep learning* berbasis arsitektur *transformer* yang dirancang untuk memahami lebih dari 100 bahasa secara bersamaan, termasuk Bahasa Indonesia [12]. mBERT dilatih menggunakan korpus besar teks *Wikipedia* dari berbagai bahasa, sehingga mampu melakukan *transfer learning* untuk berbagai tugas NLP lintas bahasa. Tidak seperti model monolingual, mBERT dapat digunakan langsung pada data berbahasa Indonesia tanpa pelatihan dari awal (*from scratch*) [13, 14].

Keunggulan utama mBERT adalah kemampuannya untuk memahami konteks kata dalam berbagai bahasa, sehingga sangat cocok untuk aplikasi NER pada teks yang menggunakan istilah campuran seperti catatan medis Indonesia. mBERT juga terbukti mampu mengenali entitas pada teks dengan distribusi kata dan tata bahasa yang berbeda-beda, termasuk variasi istilah medis, singkatan, dan penulisan diagnosis yang tidak baku [15].

Selain itu, mBERT menawarkan fleksibilitas tinggi dalam pengembangan sistem NLP untuk bahasa-bahasa yang sumber dayanya terbatas, seperti Bahasa Indonesia. Penggunaan mBERT sebagai *backbone* dalam *pipeline* ekstraksi diagnosis medis memungkinkan peneliti untuk memanfaatkan keunggulan model *pre-trained* tanpa harus membangun model dari awal, sehingga menghemat waktu dan sumber daya [13, 14].

Penelitian-penelitian terkini di Indonesia telah membuktikan bahwa mBERT memberikan performa yang unggul dalam tugas NER medis dibandingkan model konvensional. Model ini juga mudah diadaptasi untuk berbagai kasus penggunaan lain di sektor kesehatan, seperti klasifikasi topik rekam medis, ekstraksi informasi laboratorium, hingga identifikasi gejala dan obat [15, 12].

2.4 BIO Tagging

Skema BIO *tagging* adalah metode pelabelan token dalam tugas NER yang terdiri dari tiga label utama: B *(Beginning)* untuk token pertama dari entitas, I *(Inside)* untuk token yang merupakan bagian lanjutan dari entitas yang sama, dan O *(Outside)* untuk token yang bukan bagian dari

entitas apa pun [27]. Misalnya, dalam kalimat ”pasien didiagnosis diabetes mellitus tipe dua”, kata ”diabetes” diberi label B-CONDITION, sementara ”mellitus”, ”tipe”, dan ”dua” diberi label I-CONDITION.

BIO *tagging* menjadi standar *de facto* dalam dataset NERinternasional karena memudahkan model untuk mengenali batas awal dan akhir entitas, terutama untuk entitas diagnosis yang terdiri dari lebih dari satu kata [25]. Pendekatan ini mengatasi kekurangan metode label sederhana yang hanya mengklasifikasikan setiap kata sebagai entitas atau bukan entitas, sehingga mengurangi risiko fragmentasi dan ambiguitas dalam ekstraksi diagnosis medis.

Proses pelabelan dengan skema BIO sangat penting dalam pengembangan *pipeline* NER berbasis *deep learning*, karena model seperti BERT melakukan prediksi label pada setiap token dalam teks. Dengan BIO *tagging*, model dapat belajar pola urutan token dalam sebuah entitas diagnosis secara kontekstual, yang meningkatkan akurasi ekstraksi entitas pada data medis [10].

Beberapa penelitian di bidang kesehatan juga menunjukkan bahwa skema BIO *tagging* memberikan hasil yang lebih konsisten dan mudah dievaluasi, baik untuk pelatihan model maupun untuk validasi hasil ekstraksi entitas pada data medis elektronik [25, 8].

2.5 SNOMED CT

SNOMED CT (*Systematized Nomenclature of Medicine – Clinical Terms*) adalah terminologi klinis internasional yang menyediakan kode standar untuk diagnosis, prosedur, gejala, dan berbagai konsep medis lain [21]. Dengan menggunakan SNOMED CT, setiap diagnosis medis dapat direpresentasikan dalam bentuk kode unik yang berlaku secara global, sehingga memudahkan pertukaran dan integrasi data antar sistem informasi kesehatan.

Penggunaan SNOMED CT di Indonesia semakin relevan seiring dengan kebutuhan akan interoperabilitas data klinis nasional maupun internasional. Berbagai institusi kesehatan mulai mengadopsi SNOMED CT sebagai acuan utama dalam pencatatan diagnosis agar data yang dihasilkan lebih mudah dikonversi, dianalisis, dan digunakan untuk pelaporan atau penelitian epidemiologi [16]. Standarisasi diagnosis dengan SNOMED CT juga memfasilitasi proses klaim asuransi dan pelaporan penyakit ke lembaga pemerintah.

Pemetaan otomatis hasil ekstraksi diagnosis ke kode SNOMED CT merupakan tantangan tersendiri, terutama pada data berbahasa Indonesia yang banyak menggunakan istilah lokal, variasi penulisan, dan singkatan. Oleh karena itu, pengembangan *pipeline* yang mengintegrasikan NER dengan pemetaan SNOMED CT menjadi sangat penting untuk memastikan data diagnosis benar-benar valid dan siap digunakan untuk interoperabilitas.

Dengan demikian, SNOMED CT tidak hanya berperan sebagai kamus standar diagnosis medis, tetapi juga sebagai penghubung utama dalam sistem informasi kesehatan modern yang menuntut pertukaran data secara cepat, akurat, dan dapat dipercaya [21, 20].

2.6 Fast Healthcare Interoperability Resources (FHIR)

Fast Healthcare Interoperability Resources (FHIR) adalah standar global yang dikembangkan oleh HL7 untuk memfasilitasi pertukaran data kesehatan secara elektronik [7]. FHIR dirancang dengan pendekatan modular berbasis *resource*, seperti *Condition*, *Observation*, dan

MedicationRequest, sehingga setiap data klinis dapat diwakili oleh *resource* yang spesifik dan mudah dipahami oleh sistem informasi kesehatan.

Keunggulan utama FHIR terletak pada kemampuannya untuk diintegrasikan dengan teknologi web modern, karena mendukung format pertukaran data berbasis JSON dan XML, serta memanfaatkan protokol RESTful. Dengan demikian, FHIR sangat kompatibel dengan ekosistem aplikasi kesehatan digital masa kini, baik untuk pertukaran data antar rumah sakit, integrasi dengan aplikasi *mobile*, maupun pelaporan data ke pemerintah.

Dalam konteks penelitian ini, FHIR menjadi jembatan penting untuk mengonversi hasil ekstraksi diagnosis dari *pipeline* NER ke dalam format standar yang siap untuk interoperabilitas nasional maupun global. Standar FHIR juga mendorong penggunaan terminologi klinis internasional seperti SNOMED CT pada setiap *field* diagnosis, sehingga setiap *resource* CONDITION yang dihasilkan tidak hanya terstruktur, tetapi juga terstandarisasi secara terminologi [19].

Penerapan FHIR di Indonesia masih menghadapi tantangan, terutama pada proses transformasi data diagnosis dari format teks bebas ke *resource* FHIR yang benar. Oleh sebab itu, pengembangan *pipeline* yang mampu menghasilkan *resource* FHIR secara otomatis dari hasil ekstraksi diagnosis akan mempercepat integrasi sistem informasi kesehatan Indonesia ke ekosistem digital yang lebih modern dan interoperabel [16].

2.7 Penelitian Terkait

Penelitian terkait ekstraksi diagnosis medis menggunakan NLP dan NER telah banyak dilakukan, baik di tingkat internasional maupun nasional. Sari dan Azmi (2019) mengevaluasi ekstraksi informasi diagnosis pada rekam medis elektronik berbahasa Indonesia menggunakan metode *rule-based* dan CRF, dan menemukan bahwa akurasi kedua metode ini masih rendah untuk variasi istilah klinis yang kompleks [8]. Jannah dan Amalia (2020) juga menemukan tantangan serupa, terutama pada data dengan bahasa campuran dan penulisan diagnosis tidak baku [9].

Pradana et al. (2020) memperkenalkan penggunaan model BERT untuk klasifikasi diagnosis pada rekam medis Indonesia, dan menunjukkan bahwa BERT secara signifikan meningkatkan performa NER dibandingkan metode konvensional [15]. Studi ini membuktikan bahwa penggunaan *deep learning*, khususnya arsitektur *transformer* seperti BERT dan mBert, sangat relevan untuk ekstraksi diagnosis dari teks medis berbahasa Indonesia.

Di sisi lain, penelitian terkait implementasi FHIR dan SNOMED CT di Indonesia masih terbatas namun mulai berkembang. Baskara dan Pratama (2022) menunjukkan bahwa penerapan FHIR pada sistem informasi rumah sakit dapat mempercepat proses interoperabilitas, meski masih menghadapi kendala pada konversi data diagnosis ke *resource* CONDITION yang benar dan standar [16]. Santoso dan Arifin (2019) menyoroti pentingnya standarisasi terminologi diagnosis dengan SNOMED CT untuk mendukung pertukaran data lintas institusi.

Namun, sampai saat ini, masih sangat sedikit penelitian yang secara komprehensif mengembangkan *pipeline* otomatis yang mengintegrasikan NER berbasis BERT, pemetaan diagnosis ke SNOMED CT, serta konversi ke *resource* FHIR untuk data medis berbahasa Indonesia. Hal ini menjadi celah penelitian yang ingin dijawab dalam studi ini.

Tabel 2.1. Perbandingan Penelitian Terkait dengan Penelitian Ini

Peneliti	**Metode / Teknologi**	**Kelemahan**	**Kontribusi Penelitian Ini**
Sari & Azmi (2019)	Rule-based, CRF	Rendah dalam mengenali istilah tidak baku	Menggunakan mBERT untuk konteks yang lebih kompleks
Jannah & Amalia (2020)	NER CRF manual	Tidak menangani bahasa campuran	BIO Tagging otomatis dan preprocessing sistematis
Pradana et al. (2020)	BERT untuk klasifikasi diagnosis	Belum integrasi SNOMED CT/FHIR	Menambahkan pemetaan SNOMED CT dan konversi FHIR
Baskara & Pratama (2022)	FHIR untuk interoperabilitas	Belum otomatisasi diagnosis ke FHIR	Pipeline otomatis dari NER hingga output FHIR Condition
Penelitian Ini	mBERT + BIO + SNOMED CT + FHIR	–	Integrasi menyeluruh: ekstraksi diagnosis, standarisasi, dan interoperabilitas

Tabel 2.1 menampilkan ringkasan perbandingan antara penelitian-penelitian sebelumnya dan pendekatan yang dikembangkan dalam penelitian ini. Terlihat bahwa belum ada studi yang secara menyeluruh menggabungkan proses ekstraksi entitas diagnosis menggunakan mBERT, pemetaan ke terminologi SNOMED CT, serta konversi langsung ke dalam format FHIR secara otomatis untuk teks medis berbahasa Indonesia.