BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam pengembangan sistem informasi kesehatan di Indonesia, tantangan utama yang dihadapi adalah bagaimana mengekstraksi, menstandarkan, dan memvalidasi data klinis secara otomatis agar dapat mendukung pengelolaan layanan kesehatan yang efisien dan terintegrasi [1, 2, 3]. Kompleksitas bahasa Indonesia dengan segala variasi istilah medis, singkatan, dan konteks klinis yang beragam menjadikan proses ekstraksi informasi dari rekam medis elektronik sebagai tugas yang sangat menantang [4, 5]. Kebutuhan akan akurasi tinggi dalam pengolahan data klinis menjadi semakin krusial seiring dengan meningkatnya digitalisasi layanan kesehatan dan tuntutan untuk menghasilkan informasi yang dapat diandalkan untuk pengambilan keputusan klinis [6, 7].

Penelitian sebelumnya menunjukkan bahwa metode ekstraksi informasi diagnosis berbasis *rule-based* atau *Conditional Random Fields (CRF)* yang selama ini digunakan di berbagai rumah sakit Indonesia masih memiliki akurasi yang rendah dalam mengenali variasi istilah klinis serta konteks pada rekam medis elektronik berbahasa Indonesia [8, 9, 10]. Keterbatasan pendekatan konvensional ini menjadi hambatan serius dalam upaya pemanfaatan data klinis secara optimal, mengingat keragaman terminologi medis yang digunakan di berbagai institusi kesehatan di Indonesia. Variasi penulisan diagnosis, penggunaan bahasa campuran antara Indonesia dan Latin, serta inkonsistensi dalam penggunaan singkatan medis semakin mempersulit proses ekstraksi informasi yang akurat [1, 11].

Seiring berkembangnya teknologi kecerdasan buatan, penerapan model *deep learning* berbasis BERT telah terbukti secara signifikan meningkatkan performa tugas *Named Entity Recognition (NER)* pada berbagai domain, termasuk bidang medis [12, 13, 14]. Kemampuan BERT dalam memahami konteks *bidirectional* dan representasi semantik yang mendalam menjadikannya kandidat yang menjanjikan untuk mengatasi tantangan ekstraksi entitas medis pada teks berbahasa Indonesia [10, 4]. Namun, implementasi BERT secara spesifik untuk data medis lokal masih terbatas dan membutuhkan riset lebih lanjut untuk membuktikan efektivitasnya dalam menangani karakteristik unik rekam medis Indonesia, terutama dalam hal pengenalan entitas diagnosis yang kompleks dan beragam [1, 15].

Sementara itu, kebutuhan akan interoperabilitas antar sistem informasi kesehatan semakin mendesak, terutama dalam rangka mewujudkan pertukaran data klinis yang aman dan efisien di tingkat nasional. Standar global seperti *Fast Healthcare Interoperability Resources* (FHIR) dan terminologi SNOMED CT mulai diadopsi di berbagai institusi kesehatan di Indonesia, namun masih terdapat kendala dalam mengonversi data diagnosis yang beragam ke dalam format standar secara otomatis [16, 17, 18]. Padahal, standardisasi ini sangat penting agar data klinis dapat dimanfaatkan untuk analisis, pelaporan nasional, serta pengembangan inovasi digital di bidang kesehatan [7, 19, 2]. Kesenjangan antara kebutuhan interoperabilitas dan kemampuan teknis untuk mengimplementasikannya secara otomatis menjadi hambatan utama dalam realisasi ekosistem kesehatan digital yang terintegrasi [7, 20, 21].

Selain itu, proses pengisian dan pengkodean diagnosis pada rekam medis elektronik secara

manual kerap menimbulkan kesalahan, baik akibat faktor manusia maupun kurangnya pemahaman terhadap kode diagnosis yang tepat [22, 11]. Kondisi ini dapat berdampak negatif terhadap klaim asuransi, validitas data epidemiologi, dan mutu pengambilan keputusan klinis [23, 24, 3]. Kesalahan dalam pengkodean diagnosis tidak hanya mempengaruhi aspek administratif seperti penagihan dan *reimbursement*, tetapi juga dapat mengakibatkan bias dalam analisis epidemiologi dan penelitian kesehatan [11]. Oleh karena itu, pengembangan sistem otomatis berbasis kecerdasan buatan untuk melakukan audit dan validasi kode diagnosis sangat dibutuhkan guna meningkatkan akurasi, efisiensi, serta keandalan data klinis dalam rekam medis elektronik [23, 24].

Berdasarkan permasalahan tersebut, penelitian ini berupaya mengembangkan dan mengevaluasi *pipeline* otomatis berbasis model BERT untuk ekstraksi entitas klinis, pemetaan ke terminologi SNOMED CT, serta konversi ke *resource* FHIR, sehingga dapat meningkatkan mutu, interoperabilitas, dan keakuratan data klinis dalam sistem informasi kesehatan Indonesia [15, 16]. Pendekatan terintegrasi ini diharapkan dapat menjadi solusi komprehensif yang tidak hanya mengatasi tantangan ekstraksi entitas medis dengan akurasi tinggi, tetapi juga memfasilitasi standardisasi dan interoperabilitas data klinis sesuai dengan standar internasional [7, 19]. Dengan demikian, penelitian ini berkontribusi pada upaya transformasi digital sektor kesehatan Indonesia menuju ekosistem yang lebih efisien, akurat, dan terintegrasi [2, 3].

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan, maka rumusan masalah dalam skripsi yang akan dilakukan sebagai berikut:

1. Bagaimana merancang dan mengimplementasikan *pipeline* berbasis *Multilingual* BERT dan BIO *tagging* untuk melakukan identifikasi serta ekstraksi entitas diagnosis medis dari teks berbahasa Indonesia, dan memetakannya ke kode standar SNOMED CT serta format FHIR?
2. Bagaimana tingkat akurasi, *F1 Score, precision*, dan *recall* tersebut dalam melakukan identifikasi diagnosis medis pada teks berbahasa Indonesia?

1.3 Batasan Permasalahan

Berdasarkan rumusan masalah yang telah dijabarkan, maka batasan masalah dalam penelitian ini ditetapkan sebagai berikut agar penelitian lebih terfokus dan tidak meluas ke luar konteks yang telah ditentukan:

1. Penelitian ini hanya memfokuskan pada proses identifikasi dan ekstraksi entitas diagnosis medis pada teks berbahasa Indonesia, tidak mencakup entitas klinis lain seperti prosedur, obat, atau hasil laboratorium.
2. Model yang digunakan adalah *Multilingual* BERT dengan skema pelabelan BIO *tagging*, tanpa melakukan perbandingan dengan arsitektur model lain seperti LSTM, CRF, atau *rule-based*.
3. *Mapping* ke terminologi standar dilakukan hanya untuk diagnosis medis, menggunakan kode SNOMED CT sebagai referensi utama.

4. Proses konversi ke format standar interoperabilitas terbatas pada *resource* CONDITION dalam FHIR, dan tidak mencakup *resource* FHIR lainnya.
5. Evaluasi *pipeline* difokuskan pada metrik *accuracy* identifikasi diagnosis dan tidak membahas aspek performa lain seperti waktu komputasi atau kebutuhan sumber daya.
6. Data yang digunakan dalam penelitian merupakan kumpulan teks medis berbahasa Indonesia yang telah dianonimkan, bersifat sintetis atau terbuka, dan hanya mencakup data diagnosis, bukan keseluruhan rekam medis pasien yang bersifat rahasia.

1.4 Tujuan Penelitian

Berdasarkan latar belakang yang telah disampaikan, maka tujuan dalam skripsi yang akan dilakukan sebagai berikut:

1. Merancang dan mengimplementasikan *pipeline* otomatis berbasis `Multilingual` BERT dan BIO *tagging* untuk melakukan identifikasi serta ekstraksi entitas diagnosis medis dari teks berbahasa Indonesia, kemudian memetakannya ke kode standar SNOMED CT dan format FHIR.
2. Mengukur tingkat akurasi, *F1 Score, precision*, dan *recall pipeline* dalam melakukan identifikasi diagnosis medis pada teks berbahasa Indonesia.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan kontribusi dalam pengembangan metode otomatis untuk identifikasi dan ekstraksi diagnosis medis pada teks berbahasa Indonesia, sehingga dapat meningkatkan efisiensi proses pengolahan data klinis di institusi kesehatan.
2. Mendukung penerapan standarisasi terminologi medis di Indonesia melalui pemetaan otomatis diagnosis ke kode SNOMED CT, yang mempermudah integrasi dan pertukaran data antar sistem informasi kesehatan.
3. Memfasilitasi interoperabilitas data klinis secara nasional dengan mengonversi hasil identifikasi diagnosis ke dalam format FHIR, sehingga data dapat digunakan dalam berbagai aplikasi kesehatan digital.
4. Meningkatkan kualitas dan akurasi data diagnosis medis dalam rekam medis elektronik, yang dapat berdampak pada validitas pelaporan, klaim asuransi, dan pengambilan keputusan klinis.
5. Menyediakan referensi dan landasan bagi penelitian selanjutnya di bidang pemrosesan bahasa alami (NLP) untuk aplikasi medis, khususnya dalam konteks bahasa Indonesia.

1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut:

a. Bab 1 **Pendahuluan**: berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

b. Bab 2 **Tinjauan Pustaka**: membahas teori-teori dan literatur yang mendukung penelitian, seperti *Natural Language Processing* (NLP), *Named Entity Recognition* (NER), *Multilingual* BERT, skema BIO `tagging`, SNOMED CT, dan standar FHIR.

c. Bab 3 **Metodologi Penelitian**: menjelaskan metode pengumpulan data, *preprocessing* data, perancangan dan implementasi model NER berbasis *Multilingual* BERT dan BIO *tagging*, proses *mapping* ke SNOMED CT, konversi hasil ke format FHIR, serta prosedur evaluasi akurasi model.

d. Bab 4 **Hasil dan Pembahasan**: menyajikan hasil implementasi *pipeline*, evaluasi model, analisis akurasi identifikasi diagnosis, serta pembahasan temuan penelitian.

e. Bab 5 **Kesimpulan dan Saran**: berisi kesimpulan dari penelitian yang telah dilakukan serta saran untuk pengembangan atau penelitian selanjutnya.