

BAB I

PENDAHULUAN

Bab ini membahas mengenai latar belakang, rumusan masalah, tujuan dan manfaat penelitian, ruang lingkup penelitian, dan sistematika penulisan yang digunakan dalam skripsi berjudul “Optimasi Arsitektur IndoBERT-RNN-CRF untuk *Named Entity Recognition* pada Teks Medis Terkait Kesehatan Ibu dan Anak Berbahasa Indonesia”.

1.1 Latar Belakang

Kesehatan ibu dan anak merupakan salah satu indikator penting dalam pembangunan kesehatan nasional karena berkaitan langsung dengan kualitas generasi mendatang. Di Indonesia, Angka Kematian Ibu (AKI) berada di kisaran 189 per 100.000 kelahiran hidup, masih jauh dari target *Sustainable Development Goals* (SDGs) 2030 yaitu kurang dari 70 per 100.000 kelahiran hidup (Badan Pusat Statistik Indonesia, 2023; United Nations, 2015). Selain itu, prevalensi *stunting* nasional masih berada pada angka 19,8%, yang menunjukkan bahwa masalah gizi kronis pada anak belum sepenuhnya teratasi (Kementerian Kesehatan Republik Indonesia, 2025). Kondisi tersebut semakin dipengaruhi oleh ketimpangan akses layanan kesehatan dan distribusi tenaga medis antarwilayah, terutama di daerah rural dan luar Jawa–Bali, yang menyebabkan masyarakat memiliki peluang lebih rendah untuk memperoleh layanan kesehatan ibu dan anak secara berkelanjutan (Widyaningsih dkk., 2025).

Di sisi lain, perkembangan transformasi digital mendorong masyarakat untuk memanfaatkan layanan *telemedicine* dan platform konsultasi kesehatan daring sebagai sumber informasi kesehatan yang lebih mudah diakses. Mulyadita dkk. (2025) menjelaskan bahwa layanan *telemedicine* di Indonesia berkembang sangat cepat, terutama sejak pandemi COVID-19, dan menjadi salah satu alternatif utama masyarakat dalam mengakses layanan kesehatan. Salah satu platform kesehatan digital terbesar yang berkembang di Indonesia adalah Alodokter. Mulyana (2021) melaporkan bahwa Alodokter melayani sekitar 30 juta pengguna aktif setiap bulan. Tingginya interaksi pada komunitas tanya-jawab platform ini menghasilkan data teks medis dalam jumlah besar yang berpotensi menjadi sumber informasi kesehatan yang bernilai.

Data yang muncul pada forum konsultasi kesehatan daring seperti Alodokter umumnya berbentuk teks bebas yang tidak terstruktur. Hal tersebut dapat mempersulit proses ekstraksi informasi secara manual maupun menggunakan pendekatan berbasis aturan (*rule-based*) (I. Li dkk., 2021). Berbeda dengan catatan medis formal, pertanyaan pengguna pada forum kesehatan sering kali ditulis dalam bentuk narasi panjang yang memuat keluhan, riwayat singkat, serta pertanyaan lanjutan dalam satu rangkaian kalimat. Selain itu, teks yang dihasilkan pengguna juga banyak mengandung variasi penulisan, singkatan, dan bahasa tidak baku.

Kondisi tersebut mendorong perlunya pendekatan otomatis yang mampu mengenali informasi penting secara lebih cepat, konsisten, dan terstruktur. Salah satu pendekatan yang umum digunakan untuk tujuan tersebut adalah *Named Entity Recognition* (NER), yaitu teknik untuk mengidentifikasi dan mengklasifikasikan entitas penting di dalam teks ke dalam kategori tertentu. Dalam domain medis, NER berperan penting dalam membantu proses ekstraksi informasi klinis dari teks bebas sehingga dapat dimanfaatkan dalam berbagai aplikasi kesehatan, seperti pencarian informasi medis, analisis data klinis, dan sistem pendukung keputusan. Fraile Navarro dkk. (2023) menegaskan bahwa *clinical NER* merupakan elemen penting dalam mengembangkan sistem pembelajaran (*learning systems*) serta membuka potensi penuh kecerdasan buatan dari teks bebas klinis demi meningkatkan standar pelayanan kesehatan.

Dalam konteks kesehatan ibu dan anak, NER dapat digunakan untuk mengekstraksi berbagai informasi penting yang terkandung dalam teks konsultasi kesehatan, seperti gejala, penyakit, obat, tindakan medis, usia kehamilan, hasil pemeriksaan, maupun informasi fisiologis lainnya. Informasi yang semula tersimpan dalam bentuk narasi bebas dapat diubah menjadi data terstruktur sehingga lebih mudah dianalisis dan dimanfaatkan oleh sistem kesehatan digital. Sebagai contoh, sistem dapat secara otomatis mengidentifikasi keluhan yang sering dialami ibu hamil, memantau pola pertumbuhan dan perkembangan anak berdasarkan informasi pengukuran yang disebutkan pengguna, maupun mendukung proses pencarian dan pengelompokan konsultasi kesehatan berdasarkan entitas medis yang terkandung di dalamnya. Dengan demikian, NER dapat menjadi tahap awal yang penting dalam proses ekstraksi informasi kesehatan ibu dan anak dari data konsultasi daring yang tidak terstruktur.

Keberhasilan model NER dalam mengidentifikasi entitas medis sangat bergantung pada kualitas representasi kata (*word embedding*) yang digunakan. Pendekatan representasi kata tradisional seperti *static embedding* memiliki keterbatasan karena satu kata akan selalu direpresentasikan dengan vektor yang sama meskipun digunakan pada konteks kalimat yang berbeda. Untuk mengatasi keterbatasan tersebut, berkembang arsitektur *Transformer* yang memanfaatkan mekanisme *self-attention* (Vaswani dkk., 2017). Salah satu model yang dikembangkan berdasarkan arsitektur tersebut adalah *Bidirectional Encoder Representations from Transformers* (BERT), yaitu model pra-latih yang mampu memahami konteks kata dari dua arah sekaligus (Devlin dkk., 2019).

Penelitian ini memanfaatkan *Indonesian Bidirectional Encoder Representations from Transformers* (IndoBERT) sebagai *encoder* utama karena seluruh data yang dianalisis menggunakan bahasa Indonesia. IndoBERT merupakan model BERT yang dilatih menggunakan korpus bahasa Indonesia sehingga lebih sesuai untuk menangani karakteristik bahasa Indonesia dibandingkan model pra-latih berbahasa Inggris (Koto dkk., 2020). Dengan kemampuan representasi kontekstual tersebut, IndoBERT dapat memahami perbedaan makna kata berdasarkan konteks kalimat. Hal ini penting dalam teks medis karena suatu kata dapat memiliki makna berbeda tergantung penggunaannya dalam kalimat, misalnya merujuk pada gejala, kondisi tubuh, maupun tindakan medis. Oleh karena itu, penggunaan IndoBERT diharapkan mampu meningkatkan kemampuan model dalam mengenali entitas medis pada teks kesehatan ibu dan anak berbahasa Indonesia.

Meskipun IndoBERT efektif dalam membentuk representasi kontekstual melalui mekanisme *self-attention*, tugas NER merupakan tugas *sequence labeling* yang bergantung pada urutan token dan dependensi sekuensial. Hou dkk. (2024) menunjukkan bahwa penghapusan lapisan sekuensial dari arsitektur berbasis BERT menyebabkan penurunan performa karena dependensi sekuensial yang kompleks tidak tertangani secara optimal. Temuan tersebut mengindikasikan bahwa lapisan sekuensial masih memberikan kontribusi penting dalam tugas NER. Zhao dkk. (2024) juga menegaskan bahwa lapisan RNN membantu model memproses dependensi sekuensial dan memperkuat pemahaman semantik pada tingkat kalimat.

Dua arsitektur RNN yang kerap digunakan untuk tugas ini adalah *Bidirectional Long Short-Term Memory* (BiLSTM) dan *Bidirectional Gated Recurrent Unit* (BiGRU). BiLSTM dikenal kuat dalam mempelajari dependensi konteks jangka panjang dari arah maju dan

mundur melalui tiga gerbang internalnya (*forget*, *input*, dan *output gate*), tetapi konsekuensinya menuntut jumlah parameter yang besar dan pelatihan yang lebih lama (Nurlankyzy & Kulakayeva, 2025). Sebagai alternatif yang lebih efisien, BiGRU menawarkan struktur gerbang yang lebih ringkas (*update* dan *reset gate*), sehingga mampu memangkas waktu komputasi pelatihan secara signifikan dengan jumlah parameter yang jauh lebih sedikit tetapi tetap mempertahankan kualitas representasi sekuensial yang kompetitif (Qin dkk., 2021; Tong dkk., 2021).

Di samping lapisan sekuensial, *Conditional Random Field* (CRF) merupakan metode probabilistik klasik yang masih sangat relevan dan menjadi komponen fundamental dalam tugas *sequence labeling*. Sebagai salah satu pendekatan *classical* NER, CRF tidak mengklasifikasikan token secara independen, melainkan bekerja dengan memodelkan probabilitas bersyarat dari urutan label secara global (Lafferty dkk., 2001). Kemampuan ini sangat penting untuk mempelajari pola transisi label yang valid berdasarkan aturan tata bahasa, misalnya memastikan label lanjutan (*Inside*) tidak muncul tanpa didahului label awal (*Beginning*) pada skema pelabelan *Beginning, Inside, Outside* (BIO). Oleh karena itu, integrasi antara kekuatan representasi fitur sekuensial dari lapisan RNN dengan kemampuan optimalisasi transisi keputusan dari lapisan klasik CRF melahirkan kombinasi arsitektur *hybrid* yang andal dalam menjaga konsistensi hasil prediksi keputusan label secara utuh (Lample dkk., 2016).

Penerapan model *hybrid* berbasis sekuensial dan klasikal ini telah dieksplorasi dalam konteks data medis bahasa Indonesia. Penelitian Kusumawardani & Kusumawati (2024) menunjukkan bahwa arsitektur BiLSTM-CRF mampu memberikan performa yang baik pada tugas NER untuk layanan konsultasi kesehatan berbahasa Indonesia. Sejalan dengan hal tersebut, Ignasius dkk. (2025) juga berhasil membuktikan keandalan kombinasi IndoBERT-BiLSTM-CRF untuk mengekstraksi entitas dari berita kesehatan. Di sisi lain, efisiensi arsitektur berbasis BiGRU telah dibuktikan dalam berbagai penelitian internasional pada domain spesifik. Pada domain medis, kombinasi BERT-BiGRU-CRF efektif untuk mengekstraksi istilah khusus pada pengobatan tradisional Tibet dan entitas rekam medis elektronik berbahasa Mandarin (Hou dkk., 2024; Qin dkk., 2021).

Melalui eksperimen ablasi, Hou dkk. (2024) menunjukkan bahwa konfigurasi BERT-CRF maupun BiGRU-CRF mengalami penurunan *F1-score* dibandingkan arsitektur utuh BERT-BiGRU-CRF. Temuan tersebut membuktikan bahwa representasi kontekstual dari

BERT dan pemodelan sekuensial dari BiGRU saling melengkapi dalam meningkatkan kinerja tugas NER. Keunggulan efisiensi arsitektur ini juga diperkuat oleh Tong dkk. (2021) pada domain *spacecraft*, di mana model berbasis BiGRU mampu menghasilkan waktu pelatihan yang lebih singkat dibandingkan BiLSTM dengan kualitas performa yang tetap kompetitif.

Meskipun model-model standar tersebut menunjukkan hasil yang menjanjikan pada korpus formal atau rekam medis yang terstruktur, teks konsultasi kesehatan daring di Indonesia memiliki karakteristik tersendiri karena bersumber dari interaksi komunitas awam. Karakteristik teks yang memiliki variasi ekspresi yang tinggi, penggunaan bahasa tidak baku, serta derau (*noise*) penulisan menjadi tantangan tersendiri bagi model sekuensial standar dalam menangkap informasi medis secara optimal. Oleh karena itu, diperlukan eksplorasi lebih lanjut mengenai strategi modifikasi arsitektur model agar sistem lebih adaptif dalam mengenali entitas di tengah ketidakaturan teks tersebut.

Berdasarkan uraian tersebut, penelitian ini diarahkan untuk melakukan optimasi arsitektur IndoBERT-RNN-CRF melalui serangkaian strategi modifikasi struktural untuk meningkatkan kinerja pengenalan entitas medis pada domain kesehatan ibu dan anak berbahasa Indonesia. Strategi modifikasi arsitektur tersebut diimplementasikan melalui tiga tahapan pengembangan model. Tahap pertama berfokus pada peningkatan kapasitas representasi fitur melalui perluasan *hidden dimension* pada lapisan sekuensial. Ukuran *hidden state* yang lebih besar diketahui berpengaruh terhadap kemampuan model dalam mempelajari representasi sekuens yang lebih kaya (Teng dkk., 2021). Tahap kedua dilakukan melalui penerapan arsitektur bertingkat (*stacked layer*) untuk memperdalam pemodelan dependensi kontekstual dan hubungan jarak jauh antartoken, yang telah dilaporkan mampu meningkatkan performa berbagai tugas *sequence learning* (Ding dkk., 2019; L. Li dkk., 2021). Tahap ketiga dilakukan melalui optimasi lapisan klasifikasi menggunakan *dual fully connected layer* dan fungsi aktivasi *Gaussian Error Linear Unit* (GELU) untuk memperkuat transformasi representasi fitur sebelum proses prediksi label, sebagaimana telah diterapkan pada berbagai arsitektur *deep learning* modern (Hendrycks & Gimpel, 2016; Meng dkk., 2021). Melalui tahapan optimasi bertingkat ini, diharapkan diperoleh konfigurasi arsitektur yang mampu menghasilkan keseimbangan terbaik antara kualitas representasi fitur, pemodelan konteks sekuensial, dan akurasi klasifikasi entitas pada teks kesehatan ibu dan anak berbahasa Indonesia.

1.2 Rumusan Masalah

Berdasarkan latar belakang sebelumnya, rumusan masalah dalam penelitian ini adalah bagaimana mengoptimalkan arsitektur IndoBERT-RNN-CRF untuk melakukan NER pada teks medis berbahasa Indonesia. Adapun permasalahan khusus yang akan dibahas dalam penelitian ini meliputi:

1. Bagaimana performa model dasar (*baseline*) IndoBERT-BiGRU-CRF dan IndoBERT-BiLSTM-CRF sebagai acuan dalam mengevaluasi optimasi arsitektur untuk pengenalan entitas medis pada teks kesehatan ibu dan anak berbahasa Indonesia?
2. Bagaimana pengaruh modifikasi arsitektur melalui peningkatan dimensi *hidden* BiGRU, implementasi *stacked layer* BiGRU dan *dropout*, serta optimasi lapisan klasifikasi dengan GELU terhadap performa model IndoBERT-RNN-CRF?
3. Bagaimana konfigurasi arsitektur IndoBERT-RNN-CRF yang paling optimal dalam mengenali entitas medis pada teks kesehatan ibu dan anak berbahasa Indonesia?

1.3 Tujuan dan Manfaat

Berdasarkan rumusan masalah yang telah ditetapkan, tujuan penelitian ini adalah sebagai berikut:

1. Menganalisis dan membandingkan performa model dasar (*baseline*) IndoBERT-BiGRU-CRF dan IndoBERT-BiLSTM-CRF sebagai acuan dalam mengevaluasi optimasi arsitektur untuk pengenalan entitas medis pada teks kesehatan ibu dan anak berbahasa Indonesia.
2. Mengevaluasi pengaruh modifikasi arsitektur melalui peningkatan dimensi *hidden* BiGRU, implementasi *stacked layer* BiGRU dan *dropout*, serta optimasi lapisan klasifikasi dengan GELU terhadap performa model IndoBERT-RNN-CRF.
3. Menentukan konfigurasi arsitektur IndoBERT-RNN-CRF yang paling optimal dalam mengenali entitas medis pada teks kesehatan ibu dan anak berbahasa Indonesia berdasarkan hasil evaluasi performa model.

Adapun manfaat yang diharapkan dari penelitian ini adalah menjadi referensi dalam penerapan arsitektur IndoBERT-RNN-CRF pada tugas NER untuk teks medis berbahasa Indonesia, khususnya pada domain kesehatan ibu dan anak. Penelitian ini juga diharapkan dapat memberikan gambaran mengenai pengaruh berbagai strategi optimasi arsitektur terhadap performa model NER. Selain itu, hasil penelitian diharapkan dapat mendukung

pengembangan sistem ekstraksi informasi medis yang lebih akurat dan adaptif terhadap karakteristik teks konsultasi kesehatan daring yang tidak terstruktur.

1.4 Ruang Lingkup

Ruang lingkup penelitian ini dibatasi pada beberapa aspek agar pembahasan tetap terarah dan sesuai dengan tujuan penelitian, yaitu sebagai berikut.

1. Data yang digunakan dalam penelitian ini merupakan data pertanyaan medis yang diperoleh dari komunitas Alodokter, yang mencakup kategori hamil, kehamilan, ibu hamil, gangguan kehamilan, anak, dan bayi.
2. Data yang digunakan hanya pertanyaan pengguna, sedangkan bagian jawaban dokter tidak disertakan.
3. Penelitian ini dibatasi pada arsitektur IndoBERT-RNN-CRF karena kombinasi representasi kontekstual dari IndoBERT, pemodelan sekuensial oleh RNN, dan pemodelan transisi label oleh CRF menunjukkan performa yang baik pada tugas NER. Oleh karena itu, fokus penelitian ini diarahkan secara mendalam pada optimasi dan evaluasi berbagai modifikasi arsitektur tersebut tanpa membandingkannya dengan arsitektur lain.

1.5 Sistematika Penulisan

Sistematika penulisan digunakan untuk memberikan gambaran mengenai susunan dan alur pembahasan dalam laporan skripsi yang berjudul “Optimasi Arsitektur IndoBERT-RNN-CRF untuk *Named Entity Recognition* pada Teks Medis Terkait Kesehatan Ibu dan Anak Berbahasa Indonesia”. Sistematika penulisan laporan dijelaskan sebagai berikut.

BAB I PENDAHULUAN

Bab pendahuluan memuat latar belakang, rumusan masalah, tujuan dan manfaat, ruang lingkup, serta sistematika penulisan penelitian. Bab ini menjelaskan pentingnya pengembangan NER pada teks medis berbahasa Indonesia, khususnya pada domain kesehatan ibu dan anak.

BAB II

TINJAUAN PUSTAKA

Bab tinjauan pustaka memuat penelitian terdahulu (*state of the art*) dan teori-teori yang mendukung penelitian. Bab ini membahas konsep NER, *web scraping*, prapemrosesan teks, anotasi data berdasarkan skema pelabelan BIO, pengujian *intra-annotator agreement* menggunakan metrik *Cohen's kappa* dan *F1-score*, augmentasi data menggunakan *Mention Replacement* (MR) dan *Contextual Word Replacement* (CWR), arsitektur Transformer, IndoBERT, RNN, GRU, BiGRU, LSTM, BiLSTM, *fully connected layer* dengan fungsi aktivasi GELU, CRF, optimasi *hyperparameter* menggunakan *Grid Search* dan *Tree-structured Parzen Estimator* (TPE) pada *library* Optuna, serta evaluasi model NER menggunakan metrik *entity-level*.

BAB III

METODOLOGI PENELITIAN

Bab metodologi penelitian menjelaskan tahapan dan alur penelitian yang digunakan dalam pengembangan model NER berbasis IndoBERT-RNN-CRF. Tahapan penelitian meliputi pengumpulan data dari platform Alodokter, pembersihan data, anotasi entitas medis, pengujian *intra-annotator agreement*, konversi data ke format BIO, pembagian dataset, augmentasi data menggunakan MR dan CWR, pembentukan input model, pelatihan dan pengembangan model pada beberapa skenario eksperimen, pemilihan model terbaik, serta evaluasi model menggunakan metrik *entity-level*.

BAB IV

HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil eksperimen dan pembahasan dari seluruh tahapan penelitian yang telah dilakukan. Pembahasan mencakup hasil pengumpulan dan pembersihan data, hasil anotasi dan konversi BIO, hasil augmentasi data, pelatihan model IndoBERT-BiGRU-CRF dan IndoBERT-BiLSTM-CRF pada berbagai skenario eksperimen, serta

perbandingan performa antar skenario. Selain itu, dilakukan evaluasi lebih lanjut terhadap model terbaik menggunakan *Nervaluate*, *classification report*, *confusion matrix*, dan *error analysis* untuk menilai kemampuan model dalam mengenali entitas medis secara lebih komprehensif, termasuk implikasinya terhadap ekstraksi informasi kesehatan ibu dan anak.

BAB V

PENUTUP

Bab penutup berisi kesimpulan yang diperoleh berdasarkan hasil penelitian dan analisis yang telah dilakukan, serta saran untuk pengembangan penelitian selanjutnya. Kesimpulan difokuskan pada hasil optimasi arsitektur terhadap performa model IndoBERT-RNN-CRF dalam tugas NER pada teks medis berbahasa Indonesia.