BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi informasi telah mengubah cara masyarakat mengonsumsi dan mengakses informasi kesehatan, yang diperparah oleh maraknya penyebaran disinformasi berdampak serius terhadap keselamatan publik. Permasalahan ini diperburuk oleh potensi media sosial dalam menyebarkan sensasi dan ketakutan yang merusak kepercayaan terhadap sumber resmi serta profesional kesehatan [1]. Fenomena kelebihan informasi pada masa krisis kesehatan ini, yang dikenal sebagai *infodemik*, telah terbukti secara sistematis memberikan dampak buruk terhadap masyarakat, sehingga memerlukan tindakan multisektoral untuk menanggulanginya [2]. Dalam konteks lokal, narasi yang beredar termasuk dari sumber tidak terpercaya dapat bercampur dengan kepercayaan kultural dan mitos yang sudah ada, sehingga disalahpahami sebagai sains berbasis bukti [3].

Di Indonesia, penyebaran disinformasi kesehatan telah memicu kebingungan publik sekaligus memperparah situasi infodemik, dengan lebih dari 1.800 jenis hoaks dan disinformasi terkait isu kesehatan yang menjangkau 30–60% populasi [4]. Kondisi ini diperparah oleh lanskap media sosial yang luas serta rendahnya tingkat literasi kesehatan publik di Indonesia, yang turut mempercepat penyebaran misinformasi viral dan melemahkan efektivitas respons kesehatan masyarakat secara nasional [5]. Secara khusus pada isu kesehatan, jumlah disinformasi yang dilaporkan bahkan mencapai 2.256 kasus pada rentang 2018–2023 [6]. Kerentanan masyarakat terhadap disinformasi ini diperparah oleh ekosistem kepercayaan yang kompleks di ranah digital, di mana publik memberikan kepercayaan secara bersamaan terhadap sumber profesional maupun informal yang tidak terverifikasi [7]. Kondisi ini semakin mengkhawatirkan mengingat rendahnya tingkat pengetahuan pengobatan di masyarakat, yang berpotensi mendorong praktik berbahaya seperti mengonsumsi obat tidak terbukti secara ilmiah atau beralih ke pengobatan alternatif non-medis yang mengancam kesehatan fisik maupun kondisi finansial individu [3].

Untuk mengatasi keterbatasan proses *fact-checking* manual dalam mengimbangi laju penyebaran disinformasi tersebut, sejumlah penelitian telah membuktikan bahwa model berbasis BERT secara konsisten mengungguli

algoritma klasik seperti SVM, *logistic regression*, dan *naive Bayes* maupun model *deep learning* dasar dalam tugas deteksi berita palsu, sekaligus menjadi pilihan paling pragmatis ketika data berlabel terbatas [8], sebuah temuan yang dikonfirmasi pula oleh tinjauan literatur lebih luas terhadap deteksi *fake news* berbasis *machine learning* dan *deep learning* [9]. Pada domain kesehatan secara spesifik, kerangka kerja berbasis BERT dan Bi-LSTM yang dioptimalkan dengan Optuna telah terbukti efektif mendeteksi disinformasi kesehatan terkait COVID-19 dengan akurasi hingga 99,47% [10], sementara pendekatan berbasis *ensemble* BlueBERT-FFNN turut dikembangkan untuk *fact-checking* artikel kesehatan berbahasa Inggris [11] dan arsitektur CNN-LSTM berbasis Word2Vec dieksplorasi untuk deteksi disinformasi COVID-19 [12], namun studi pemetaan sistematis terhadap 63 penelitian AI penangkal misinformasi kesehatan periode 2017–2022 menemukan bahwa mayoritas studi tersebut berasal dari wilayah Amerika dan Eropa sehingga kurang merepresentasikan konteks Asia Tenggara [13]. Penelitian terdahulu telah mengembangkan model *hybrid* IndoBERT-LSTM untuk deteksi hoax berbahasa Indonesia dengan meraih akurasi sebesar 93,2% [14]. Meskipun menunjukkan kinerja superior dibandingkan model pembanding, penelitian tersebut mengidentifikasi keterbatasan temporal pada model, di mana efektivitas deteksi berpotensi menurun seiring perubahan tren dan pola informasi setelah Februari 2023 [14]. Keterbatasan adaptasi terhadap dinamika informasi ini menggarisbawahi kebutuhan akan pengembangan model bahasa Indonesia yang lebih mutakhir dan *robust*, khususnya untuk mendeteksi disinformasi pada domain spesifik seperti kesehatan yang memiliki dampak kritis terhadap keselamatan publik.

Dalam perkembangan arsitektur BERT [15], terdapat sejumlah varian yang telah dikembangkan khusus untuk data berbahasa Indonesia, seperti IndoBERT, IndoRoBERTa, dan varian lainnya yang dikembangkan oleh tim IndoNLU, yang diadaptasi untuk meningkatkan efisiensi dan akurasi pemrosesan teks berbahasa Indonesia pada berbagai tugas NLP [16]. IndoBERT sendiri telah terbukti mencapai akurasi 89,50% dalam klasifikasi pesan terkait COVID-19 berbahasa Indonesia, mengungguli model BERT multilingual yang hanya meraih 81,50% [17]. Temuan ini mengonfirmasi keunggulan model yang dipre-latih khusus pada korpus bahasa Indonesia dibandingkan model umum lintas bahasa, sejalan dengan temuan yang lebih luas bahwa model monolingual yang dilatih dengan korpus memadai secara konsisten menunjukkan performa lebih baik dibandingkan model multibahasa, karena perluasan cakupan ke banyak bahasa sekaligus cenderung mengurangi kapasitas representasi khusus untuk satu bahasa tertentu [18]. Temuan

ini juga sejalan dengan penelitian lain yang menemukan bahwa IndoRoBERTa, yang mengadopsi strategi pelatihan RoBERTa dengan menghilangkan *next sentence prediction* dan menerapkan *dynamic masking* [19], menghasilkan performa lebih baik dibandingkan MBERT pada tugas klasifikasi kepribadian berbahasa Indonesia sehingga terpilih sebagai salah satu model dalam skema *ensemble learning* [20]. Namun, eksplorasi komparatif antara ketiga varian Transformer lokal ini, yaitu IndoBERT, IndoRoBERTa, dan IndoDistilBERT yang mengadopsi teknik *knowledge distillation* untuk efisiensi komputasi [21], secara khusus pada domain disinformasi kesehatan belum pernah dilakukan secara sistematis.

Pemilihan ketiga varian tersebut juga bukan tanpa pertimbangan terhadap alternatif model bahasa Indonesia lain yang tersedia. Model regional seperti NusaBERT, yang memperluas kosakata IndoBERT dengan korpus multibahasa daerah untuk mencakup keberagaman bahasa di Indonesia, dirancang khusus untuk kebutuhan multibahasa dan multikultural yang berbeda dari fokus penelitian ini pada bahasa Indonesia baku [22], sehingga perluasan cakupan bahasa tersebut, sebagaimana dijelaskan sebelumnya [18], berpotensi kurang optimal dibandingkan model yang murni dilatih untuk bahasa Indonesia standar. Sementara itu, IndoBERTweet dikembangkan melalui inisialisasi kosakata khusus yang disesuaikan dengan karakteristik bahasa informal pada platform Twitter [23], sehingga kurang sesuai untuk data penelitian ini yang berupa artikel berita dan narasi *fact-checking* berbahasa formal. Atas dasar pertimbangan tersebut, penelitian ini difokuskan pada perbandingan IndoBERT, IndoRoBERTa, dan IndoDistilBERT sebagai representasi arsitektur dasar, arsitektur yang dioptimalkan dari sisi strategi pelatihan, dan arsitektur yang dikompresi untuk efisiensi komputasi di dalam keluarga Transformer monolingual bahasa Indonesia baku.

Penelitian ini akan mengimplementasikan dan membandingkan tiga arsitektur Transformer berbahasa Indonesia, yaitu IndoBERT, IndoRoBERTa, dan IndoDistilBERT, untuk klasifikasi disinformasi pada artikel kesehatan. Selain komparasi arsitektur, penelitian ini juga melakukan analisis interpretabilitas menggunakan LIME dan teknik *text highlighting* untuk mengungkap kata-kata kunci yang mempengaruhi keputusan klasifikasi pada masing-masing model. Evaluasi kinerja menggunakan metrik akurasi, presisi, *recall*, dan *F1-score* dilakukan sebagai pelengkap, dengan penekanan pada kemampuan model dalam meminimalkan *false negative* yaitu kesalahan klasifikasi paling berisiko dalam konteks disinformasi kesehatan.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana karakteristik dataset disinformasi kesehatan memengaruhi performa masing-masing model Transformer yang diuji?
2. Bagaimana performa dan efisiensi komputasi arsitektur IndoBERT, IndoRoBERTa, dan IndoDistilBERT dalam mengklasifikasikan disinformasi kesehatan berbahasa Indonesia?
3. Bagaimana hasil interpretasi model Transformer (IndoBERT, IndoRoBERTa, dan IndoDistilBERT) melalui analisis LIME dan teknik text highlighting dalam menjelaskan kata-kata kunci yang memengaruhi keputusan klasifikasi disinformasi kesehatan berbahasa Indonesia?

1.3 Batasan Permasalahan

Batasan masalah dalam penelitian ini adalah:

1. Data yang digunakan adalah artikel kesehatan berbahasa Indonesia yang berasal dari portal berita kesehatan resmi (sumber fakta) serta portal verifikasi fakta Mafindo/TurnBackhoax dan dataset Komdigi (sumber disinformasi).
2. Klasifikasi dilakukan secara biner: Disinformasi dan Fakta.
3. Analisis interpretabilitas menggunakan metode LIME yang bersifat lokal (per sampel) dan tidak merepresentasikan perilaku model secara global.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Menganalisis pengaruh karakteristik dataset terhadap performa masing-masing model Transformer yang diuji.
2. Melakukan komparasi performa dan efisiensi komputasi antara model IndoBERT, IndoRoBERTa, dan IndoDistilBERT dalam mengklasifikasikan disinformasi kesehatan berbahasa Indonesia.

3. Melakukan analisis interpretabilitas model menggunakan LIME dan teknik text highlighting untuk mengidentifikasi kata-kata kunci yang memengaruhi keputusan klasifikasi disinformasi kesehatan berbahasa Indonesia.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Penelitian ini memberikan wawasan mengenai pentingnya karakteristik dataset, seperti distribusi kelas dan variasi panjang teks, dalam pengembangan dataset deteksi disinformasi kesehatan berbahasa Indonesia. Hasil penelitian ini diharapkan dapat menjadi acuan bagi penelitian selanjutnya dalam menyusun dataset yang lebih representatif sehingga dapat meminimalkan potensi bias pada proses pelatihan model.

2. Penelitian ini memberikan acuan dalam membandingkan performa dan efisiensi komputasi model IndoBERT, IndoRoBERTa, dan IndoDistilBERT sehingga dapat menjadi pertimbangan bagi peneliti maupun pengembang sistem deteksi disinformasi kesehatan dalam memilih arsitektur Transformer yang sesuai dengan kebutuhan dan keterbatasan sumber daya komputasi.

3. Penelitian ini memberikan wawasan mengenai interpretabilitas model deteksi disinformasi kesehatan melalui analisis LIME dan teknik *text highlighting*, sehingga dapat membantu memahami kata-kata yang berkontribusi terhadap keputusan klasifikasi serta meningkatkan transparansi hasil prediksi model.

1.6 Sistematika Penulisan

Sistematika penulisan laporan penelitian ini terdiri atas lima bab sebagai berikut.

1. Bab 1 PENDAHULUAN: menjelaskan latar belakang disinformasi kesehatan, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.
2. Bab 2 LANDASAN TEORI: menyajikan landasan teori mengenai disinformasi kesehatan, arsitektur Transformer beserta mekanisme *Self-Attention* dan *Multi-Head Attention*, model IndoBERT, IndoRoBERTa, dan

IndoDistilBERT, teknik augmentasi data (*Easy Data Augmentation* dan *back-translation*), metrik evaluasi klasifikasi, serta metode *explainability* menggunakan LIME.

3. Bab 3 METODOLOGI PENELITIAN: menguraikan tahapan penelitian yang meliputi pengumpulan dataset, *preprocessing*, pembagian data sebelum augmentasi, augmentasi data, perhitungan *inverse-frequency class weights*, proses *fine-tuning* ketiga model menggunakan hasil *grid search*, serta metode evaluasi dan analisis interpretabilitas menggunakan LIME.
4. Bab 4 HASIL DAN DISKUSI: memaparkan hasil analisis karakteristik dataset, perbandingan performa ketiga model berdasarkan berbagai metrik evaluasi, analisis proses pelatihan, pengujian generalisasi, analisis interpretabilitas menggunakan LIME, serta pembahasan kelebihan dan keterbatasan masing-masing model.
5. Bab 5 SIMPULAN DAN SARAN: menyajikan simpulan berdasarkan hasil penelitian serta saran untuk pengembangan penelitian selanjutnya.