

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan pesat teknologi *digital* dan akses internet dalam beberapa tahun terakhir mengubah cara masyarakat mencari dan menerima informasi. Saat ini, beragam berita dan sumber informasi mudah dijangkau melalui saluran *digital*, sehingga internet menjadi rujukan utama bagi kebutuhan informasi banyak orang. Munculnya banyak portal berita *daring*, platform *digital*, dan media sosial mempercepat arus dalam penyebaran informasi sehingga dapat diakses secara *real-time* oleh siapa saja. Kondisi ini membawa manfaat karena publik dapat memperoleh informasi dengan lebih praktis dan efisien dibandingkan sebelumnya. Namun, kemudahan akses tersebut juga menimbulkan tantangan baru, yakni meningkatnya peredaran berita palsu atau hoaks di ruang *digital* [1],[2],[3]. Hoaks didefinisikan sebagai informasi yang sengaja dibuat atau dimanipulasi untuk menyesatkan publik dan membentuk opini tertentu [4].

Pertumbuhan jumlah pengguna internet setiap tahun memperbesar peluang tersebarnya berita palsu, terutama lewat media sosial yang memungkinkan informasi menyebar dengan sangat cepat dalam waktu singkat. Masih banyak pengguna yang menerima dan membagikan konten tanpa memverifikasi sumber atau kebenaran isi berita tersebut [5]. Situasi ini berpotensi mempengaruhi persepsi publik terhadap suatu isu dan bahkan memicu konflik sosial akibat informasi yang tidak *valid*. Rama *et al.* menyatakan bahwa dampak penyebaran hoaks tidak hanya berupa kesalahan informasi, tetapi juga berisiko mengancam stabilitas sosial dan persatuan masyarakat [6]. Oleh karena itu, diperlukan sebuah sistem yang mampu mengidentifikasi berita hoaks secara otomatis, cepat, dan dengan tingkat akurasi tinggi.

Sejumlah penelitian *terdahulu* telah mengimplementasikan metode *machine learning* maupun *deep learning* dalam mendeteksi berita hoaks, termasuk penggunaan algoritma XGBoost dan IndoBERT. Model *transformer* seperti

IndoBERT dikenal efektif dalam memahami konteks bahasa Indonesia sehingga mampu menghasilkan performa klasifikasi yang cukup baik. Di sisi lain, XGBoost juga banyak digunakan pada klasifikasi teks karena memiliki kemampuan komputasi yang efisien serta performa yang stabil. Walaupun demikian, hasil dari kedua pendekatan tersebut masih menunjukkan variasi performa pada dataset yang berbeda, sehingga tingkat akurasi belum sepenuhnya konsisten maupun optimal dalam berbagai kondisi data. Selain itu, penelitian yang mengkombinasikan IndoBERT dan XGBoost dalam bentuk model *hybrid* untuk klasifikasi berita hoaks berbahasa Indonesia masih tergolong terbatas. Hal ini menunjukkan adanya peluang untuk mengembangkan model yang dapat memanfaatkan keunggulan *contextual embedding* dari IndoBERT dan kemampuan klasifikasi milik XGBoost agar menghasilkan performa yang lebih baik serta stabil.

Berdasarkan latar tersebut, penelitian ini mengembangkan dan membandingkan tiga pendekatan klasifikasi, yakni IndoBERT, XGBoost, serta model *hybrid* yang menggabungkan kedua metode untuk mendeteksi hoaks berbahasa Indonesia. Pemilihan IndoBERT didasari oleh kemampuannya menangkap konteks dan makna bahasa secara mendalam melalui pendekatan *contextual embedding* [7]. Sedangkan XGBoost dipilih karena performa klasifikasi yang andal, efisiensi komputasi, serta kemampuan menangani data berdimensi besar dengan risiko *overfitting* relatif rendah [8],[9]. Pada model *hybrid*, *embedding* yang dihasilkan IndoBERT difungsikan sebagai fitur input bagi XGBoost sehingga kedua pendekatan saling melengkapi dalam proses klasifikasi.

Penelitian ini bertujuan memperoleh model klasifikasi dengan performa *optimal* dan akurasi lebih tinggi melalui penggabungan lima dataset dari AntaraNews, Detik, Kompas, TurnBackHoax, dan Komdigi. Kombinasi dataset diharapkan meningkatkan kemampuan generalisasi model dalam mendeteksi hoaks berbahasa Indonesia secara lebih andal dan konsisten. Selain dataset utama, penelitian juga memakai data berita tambahan sebagai *external test* untuk menguji generalisasi model terhadap berita baru yang berada di luar data latih. Pengujian ini dilakukan untuk menilai konsistensi model dalam mengenali pola hoaks maupun non-hoaks

dalam kondisi yang lebih mendekati implementasi nyata di media *digital* Indonesia. Dengan pendekatan ini diharapkan dihasilkan model deteksi hoaks yang akurat, stabil, dan memiliki kemampuan generalisasi baik pada berbagai sumber berita digital berbahasa Indonesia, serta memberi kontribusi pada pengembangan penelitian NLP dan klasifikasi teks untuk Bahasa Indonesia.

1.2 Rumusan Masalah

1. Bagaimana pengaruh keberagaman lima kombinasi dataset yang berasal dari AntaraNews, Detik, Kompas, TurnBackHoax, dan Komdigi terhadap performa model klasifikasi dalam mendeteksi berita hoaks dan non-hoaks pada media digital berbahasa Indonesia?
2. Bagaimana hasil perbandingan performa model XGBoost dan IndoBERT dalam mengklasifikasikan berita hoaks dan non-hoaks pada media digital berbahasa Indonesia berdasarkan lima kombinasi dataset yang digunakan?
3. Bagaimana performa model *hybrid* yang menggabungkan *contextual embedding* IndoBERT dan algoritma XGBoost dalam mengklasifikasikan berita hoaks dan non-hoaks pada media *digital* berbahasa Indonesia berdasarkan kombinasi dataset yang digunakan serta pengujian menggunakan *external testing data*?

1.3 Batasan Masalah

1. Penelitian ini difokuskan pada pengembangan dan evaluasi model klasifikasi untuk membedakan berita Hoaks dan Non-Hoaks pada media *digital* berbahasa Indonesia. Dataset penelitian diperoleh dari empat portal berita *daring*, yaitu AntaraNews, Detik, dan Kompas sebagai representasi berita Non-Hoaks, sedangkan TurnBackHoax digunakan sebagai sumber data Hoaks karena merupakan platform pemeriksa fakta resmi di Indonesia. Selain itu, data hoaks juga diperoleh dari portal pengecekan hoaks Komdigi yang menyediakan informasi mengenai konten dan berita yang telah teridentifikasi sebagai hoaks berdasarkan proses verifikasi yang dilakukan oleh instansi terkait. Proses klasifikasi hanya memanfaatkan konten teks artikel berita sebagai fitur utama analisis. Oleh karena itu, penelitian ini tidak mencakup proses verifikasi fakta secara langsung terhadap sumber

berita, investigasi lapangan, maupun penilaian manual oleh pakar di luar label yang telah tersedia pada dataset.

2. Ruang lingkup data penelitian dibatasi pada artikel berita *digital* berbahasa Indonesia yang bersumber dari repositori Kaggle dan hasil *web scraping*. Dataset utama untuk pelatihan dan evaluasi *internal* berjumlah 27.479 artikel yang dikumpulkan pada periode November–Desember 2025. Selain itu, digunakan pula data hasil *web scraping* sebagai data uji eksternal untuk menilai kemampuan generalisasi model terhadap berita baru di luar dataset pelatihan. Semua data yang dipakai hanya berbahasa Indonesia, sehingga penelitian ini tidak mencakup analisis terhadap berita berbahasa asing. Penelitian ini juga tidak menyelidiki perubahan pola penyebaran hoaks dalam rentang waktu panjang atau studi *longitudinal*.
3. Fokus penelitian terbatas pada penerapan dan perbandingan tiga pendekatan klasifikasi, yakni XGBoost, IndoBERT, dan model *hybrid* yang memanfaatkan *embedding* IndoBERT sebagai fitur input untuk XGBoost. Algoritma klasifikasi lain di luar ketiga metode tersebut tidak dianalisis. Optimasi *hyperparameter* dilakukan menggunakan *GridSearchCV* dengan ruang parameter yang dibatasi untuk menemukan konfigurasi model yang *optimal*. Evaluasi performa memakai confusion matrix dan metrik turunannya, yaitu *accuracy*, *precision*, *recall*, *F1-score*, dan ROC-AUC, sehingga kinerja masing-masing model dalam mengidentifikasi berita hoaks dan non-hoaks dapat dinilai secara objektif dan terukur.

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan Penelitian

1. Menganalisis pengaruh keberagaman kombinasi dataset terhadap performa model klasifikasi

Penelitian ini bertujuan untuk menganalisis pengaruh keberagaman kombinasi dataset terhadap kinerja model klasifikasi dalam mendeteksi berita Hoaks dan Non-Hoaks berbahasa Indonesia. Proses pelatihan dan evaluasi dilakukan pada empat kombinasi dataset, yaitu AntaraNews–TurnBackHoax, Detik–TurnBackHoax, Kompas–TurnBackHoax, serta

gabungan seluruh dataset. Melalui pengujian tersebut, penelitian ini menganalisis apakah perbedaan karakteristik dan komposisi data dari masing-masing sumber berita memberikan pengaruh terhadap hasil klasifikasi model XGBoost, IndoBERT, dan model *hybrid* berdasarkan metrik *accuracy*, *precision*, *recall*, dan *F1-score*.

2. Membandingkan performa model XGBoost dan IndoBERT dalam klasifikasi berita hoaks

Penelitian ini juga memiliki tujuan untuk membandingkan performa dua metode klasifikasi yang berbeda, yaitu XGBoost dan IndoBERT, dalam proses identifikasi berita hoaks dan non-hoaks pada media *digital* berbahasa Indonesia. XGBoost menggunakan representasi fitur berbasis TF-IDF, sedangkan IndoBERT menerapkan pendekatan *contextual embedding* berbasis *transformer* yang mampu memahami konteks bahasa secara lebih mendalam. Perbandingan kedua model dilakukan melalui evaluasi menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* pada seluruh kombinasi dataset, sehingga dapat diketahui model yang memiliki kemampuan klasifikasi yang lebih *optimal* dalam mendeteksi berita hoaks.

3. Mengevaluasi performa model *hybrid* berbasis IndoBERT dan XGBoost

Selain itu, penelitian ini bertujuan untuk mengembangkan serta mengevaluasi model *hybrid* yang mengkombinasikan *contextual embedding* dari IndoBERT dengan metode klasifikasi berbasis *gradient boosting* milik XGBoost. Pendekatan tersebut diterapkan untuk menganalisis apakah penggabungan kedua metode mampu menghasilkan performa klasifikasi yang lebih baik dibandingkan penggunaan XGBoost maupun IndoBERT secara terpisah. Evaluasi performa model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* guna mengukur kemampuan model *hybrid* dalam mendeteksi berita hoaks dan non-hoaks pada media *digital* berbahasa Indonesia.

1.4.2 Manfaat Penelitian

1. Manfaat Praktis

Secara praktis, penelitian ini diharapkan memberikan manfaat bagi masyarakat sebagai pengguna media *digital*. Penelitian ini menghasilkan sebuah aplikasi berbasis *web* menggunakan Streamlit yang memungkinkan pengguna melakukan deteksi berita hoaks hanya dengan memasukkan teks berita ke dalam sistem. Kehadiran aplikasi tersebut mempermudah proses identifikasi berita palsu karena dapat dilakukan secara otomatis, cepat, dan mudah digunakan tanpa membutuhkan kemampuan teknis yang mendalam.

2. Manfaat Akademis

Dari sisi akademis, penelitian ini memberikan kontribusi terhadap pengembangan metode klasifikasi teks berbahasa Indonesia, khususnya dalam bidang deteksi berita hoaks pada media *digital*. Kontribusi pertama penelitian ini terletak pada penyajian hasil perbandingan empiris antara model XGBoost berbasis TF-IDF, model IndoBERT berbasis *transformer* yang telah melalui tahap *fine-tuning*, serta model *hybrid* yang mengombinasikan *embedding* IndoBERT dengan XGBoost. Ketiga pendekatan tersebut diuji menggunakan lima kombinasi dataset yang berasal dari AntaraNews, Detik, Kompas, TurnBackHoax, dan Komdigi. Kontribusi berikutnya adalah penyediaan temuan terkait pengaruh variasi sumber data terhadap performa model klasifikasi. Hasil tersebut dapat dijadikan referensi bagi penelitian selanjutnya dalam menentukan komposisi dataset yang paling efektif untuk pengembangan sistem deteksi hoaks berbahasa Indonesia. Selain itu, penelitian ini juga menunjukkan bahwa pendekatan *hybrid* memiliki potensi dalam meningkatkan performa klasifikasi dibandingkan penggunaan model tunggal. Temuan tersebut diharapkan dapat menjadi dasar pengembangan penelitian berikutnya yang menggabungkan pendekatan *transformer* dan *machine learning* pada pemrosesan teks berbahasa Indonesia dalam cakupan yang lebih luas.

1.5 Sistematika Penulisan

BAB 1: PENDAHULUAN

Bab ini menguraikan latar belakang penelitian yang muncul dari peningkatan penyebaran hoaks di media *digital* dan kebutuhan akan sistem otomatis untuk mengidentifikasi informasi palsu secara lebih akurat. Dijelaskan pula permasalahan utama yang menjadi fokus penelitian, khususnya terkait klasifikasi berita hoaks dan non-hoaks menggunakan pendekatan *machine learning* dan *deep learning*. Selain itu disajikan rumusan masalah, tujuan penelitian, kontribusi akademis dan praktis, serta batasan penelitian untuk menjaga fokus pembahasan. Bab ditutup dengan sistematika penulisan sebagai peta isi laporan.

BAB 2: LANDASAN TEORI

Bab ini memaparkan tinjauan pustaka dan teori yang mendasari penelitian. Dimulai dengan ulasan studi terdahulu yang relevan untuk menempatkan penelitian ini dalam konteks karya sebelumnya. Dilanjutkan dengan penjelasan konsep-konsep kunci seperti deteksi hoaks, klasifikasi teks, *machine learning*, dan *deep learning*. Bab ini juga menguraikan algoritma yang digunakan (XGBoost, IndoBERT, dan model *hybrid*), teknik representasi teks, langkah-langkah *preprocessing*, serta perangkat lunak dan *library* pendukung.

BAB 3: METODOLOGI

Bab ini menjelaskan langkah-langkah penelitian secara sistematis dengan kerangka CRISP-DM. Termasuk definisi objek penelitian (deteksi hoaks berbasis teks), proses pengumpulan data, pemahaman data, dan persiapan data (meliputi pembersihan dan *text preprocessing*). Pada bagian pemodelan dijabarkan pembangunan model XGBoost, IndoBERT, dan model *hybrid* serta prosedur *hyperparameter tuning*. Bab ini juga menguraikan metode evaluasi (*confusion matrix* dan metrik turunannya) dan penggunaan data uji eksternal dari *web scraping*. Akhir bab menjelaskan proses *deployment* model terpilih ke aplikasi *web* berbasis Streamlit.

BAB 4: ANALISIS DAN HASIL PENELITIAN

Bab ini memaparkan implementasi dan analisis hasil penelitian secara mendetail. Dimulai dari hasil pengolahan data (tahap *cleaning*, *text preprocessing*, sampai dengan data siap digunakan pada proses pemodelan), kemudian proses pelatihan dan pengujian model pada tiap kombinasi dataset. Performa model dianalisis menggunakan metrik seperti *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*, disertai *classification report*, *confusion matrix*, *ROC curve*, dan *learning curve*. Selain itu disajikan hasil pengujian eksternal dari data *web scraping* dan implementasi model terbaik pada aplikasi *web* deteksi hoaks.

BAB 5: SIMPULAN DAN SARAN

Bab penutup berisi simpulan yang merangkum temuan penelitian dan menjawab rumusan masalah, termasuk perbandingan performa XGBoost, IndoBERT, dan model *hybrid* dalam mendeteksi hoaks. Dijelaskan pula kombinasi model-dataset yang memberikan hasil terbaik berdasarkan evaluasi dan pengujian eksternal. Bagian saran memuat rekomendasi untuk penelitian lanjutan, seperti peningkatan performa model, perluasan dan dinamika dataset, serta pengembangan sistem deteksi hoaks untuk implementasi yang lebih efektif di masa depan.