BAB I
PENDAHULUAN

Bab ini membahas dasar penelitian yang meliputi latar belakang permasalahan, tujuan yang ingin dicapai, serta manfaat yang diharapkan. Fokus penelitian diarahkan pada pengembangan sistem klasifikasi dan rekomendasi dokumen ilmiah berbasis teks dengan memanfaatkan model BERT yang dioptimalkan melalui proses fine-tuning, sehingga penelitian ini memiliki landasan konseptual sekaligus arah kontribusi yang jelas.

1.1 Latar Belakang

Perkembangan teknologi informasi mengalami kemajuan pesat, terutama dengan hadirnya kecerdasan buatan yang mendorong transformasi signifikan dalam berbagai bidang seperti pendidikan, kesehatan, dan ekonomi. Penerapan kecerdasan buatan mampu mempercepat otomatisasi serta mendukung pengambilan keputusan berbasis data secara efisien dan kompleks (Adiman dkk., 2024). Seiring berjalannya waktu, pertumbuhan data digital dalam bentuk teks meningkat drastis. Teks sebagai informasi tak terstruktur menimbulkan tantangan dalam pengelolaan, khususnya dalam proses pencarian dan klasifikasi informasi secara tepat (Fitra dkk., 2025).

Dalam praktik sehari-hari, peneliti maupun mahasiswa sering menghadapi kesulitan dalam menelusuri literatur ilmiah karena jumlah publikasi yang terus meningkat setiap tahun (Zuluaga dkk., 2022). Proses pencarian dan klasifikasi dokumen masih banyak dilakukan secara manual, sehingga rawan menimbulkan kebingungan, keterlambatan, dan tidak konsisten dalam menentukan referensi yang relevan (Rehman dkk., 2022). Permasalahan tersebut semakin nyata ketika sebuah topik penelitian memiliki istilah yang bersifat umum atau digunakan lintas bidang, sehingga dokumen yang ditemukan sering kali bercampur dengan topik lain yang tidak sepenuhnya sesuai.

Untuk menjawab tantangan tersebut, pendekatan *Natural Language Processing* (NLP) digunakan sebagai solusi. NLP merupakan bidang interdisipliner antara ilmu komputer, linguistik, dan kecerdasan buatan yang memungkinkan komputer memahami, memproses, dan menganalisis bahasa manusia secara

otomatis (Abdurrohim & Rahman, 2024). Salah satu model NLP yang populer adalah *Bidirectional Encoder Representations from Transformers* (BERT), model berbasis arsitektur *transformer* yang dilatih menggunakan data berskala besar dan mampu memahami makna kata dalam konteks kalimat secara dua arah. Model ini terbukti unggul dalam berbagai tugas NLP seperti klasifikasi teks, analisis sentimen, dan pencarian informasi (Sukmawati dkk., 2024).

Sistem klasifikasi teks yang banyak digunakan saat ini umumnya masih mengandalkan metode *machine learning* tradisional seperti *Naive Bayes*, *Support Vector Machine* yang biasanya dikombinasikan dengan representasi *Term Frequency–Inverse Document Frequency* (TF-IDF). Meskipun efektif untuk kasus tertentu, pendekatan ini belum mampu menangkap makna kontekstual karena hanya memanfaatkan frekuensi kata tanpa memperhatikan hubungan antarkata dalam kalimat (Haeruddin dkk., 2025). Sebaliknya, BERT menawarkan pendekatan berbasis representasi kontekstual melalui arsitektur *transformer* dan mekanisme *self-attention* yang memungkinkan pemahaman hubungan antar kata secara dua arah (Pradipta & Widodo, 2024). Hal ini menjadikan BERT sangat efektif untuk tugas NLP yang memerlukan pemahaman makna teks secara mendalam.

Selain kebutuhan klasifikasi, pengguna juga sering memerlukan sistem yang mampu merekomendasikan informasi lain yang relevan secara semantik. Dalam konteks pengelolaan informasi berbasis teks, penggabungan antara klasifikasi dan sistem rekomendasi menjadi penting untuk mendukung penelusuran dan pemanfaatan informasi secara kontekstual. Namun, integrasi kedua pendekatan ini ke dalam satu sistem masih jarang ditemukan terutama untuk data teks.

Penelitian ini mengembangkan sistem klasifikasi dan rekomendasi dokumen berbasis teks menggunakan model BERT yang telah disesuaikan melalui proses *fine-tuning*. *Fine-tuning* merupakan pelatihan lanjutan terhadap model BERT yang sebelumnya melalui *pre-training* pada data umum, kemudian diadaptasi menggunakan data domain spesifik agar dapat memahami karakteristik linguistik sesuai konteks tugas tertentu (Hassan dkk., 2022). Representasi teks yang dihasilkan dari proses tersebut digunakan untuk menentukan kategori dokumen sekaligus mengukur kemiripan semantik antar dokumen dalam ruang vektor

menggunakan metode *cosine similarity*. Metode *cosine similarity* menghitung kemiripan antara dua vektor berdasarkan sudut di antara keduanya dan menghasilkan nilai antara 0 hingga 1, di mana nilai mendekati 1 menunjukkan bahwa dua dokumen memiliki makna yang sangat mirip. Pendekatan ini efektif untuk menangkap kesamaan semantik antar teks tanpa terpengaruh oleh perbedaan panjang dokumen (Telaumbanua & Nababan, 2022).

Sistem yang dikembangkan memiliki dua fungsi utama, yaitu klasifikasi kategori laboratorium dan rekomendasi dokumen karya ilmiah. Model yang digunakan adalah *bert-base-uncased* dari *Hugging Face*, termasuk *tokenizer* yang diambil langsung dari model yang sama agar selaras dalam pemrosesan teks. Model ini disesuaikan melalui proses *fine-tuning* menggunakan data ilmiah agar lebih relevan dengan konteks bahasa domain. Dataset dibagi menjadi data *training* dan *validation* untuk memantau performa model selama pelatihan. Pelatihan dilakukan dengan menggunakan arsitektur *bert-base-uncased* untuk memastikan konsistensi antara model dan *tokenizer*. Evaluasi dilakukan dengan mengamati nilai *loss*, *accuracy*, dan F1-score pada data *validation* guna memastikan model tidak mengalami *overfitting* dan mampu melakukan generalisasi. Jika performa pada data *validation* stabil, evaluasi dilanjutkan menggunakan data testing dengan metrik *accuracy*, *precision*, *recall*, dan F1-score. Visualisasi *confusion matrix* digunakan untuk menganalisis distribusi prediksi kategori laboratorium secara lebih rinci.

Model hasil *fine-tuning* kemudian digunakan kembali dalam sistem rekomendasi tanpa pelatihan ulang. Representasi vektor dari dokumen dimanfaatkan untuk menghitung kedekatan semantik antar dokumen menggunakan metode *cosine similarity*. Evaluasi terhadap hasil rekomendasi dilakukan dengan menggunakan *Precision*@K dan *Recall*@k berdasarkan kategori laboratorium, untuk menilai relevansi dokumen yang disarankan terhadap konteks klasifikasi.

Penelitian ini difokuskan pada dokumen ilmiah di bidang ilmu komputer dan informatika dengan total 2.000 entri yang mencakup judul, abstrak, dan kata kunci sebagai fitur utama. Model yang digunakan adalah *bert-base-uncased* dari *Hugging Face* yang disesuaikan melalui proses *fine-tuning* untuk menghasilkan representasi vektor semantik sebagai dasar klasifikasi kategori laboratorium dan rekomendasi

dokumen yang memiliki kesamaan makna. Evaluasi penelitian dilakukan dengan metrik akurasi, presisi, *recall*, F1-score serta relevansi rekomendasi menggunakan *Precision*@K dan *Recall*@K, dengan batasan pada ruang lingkup tersebut sehingga hasil penelitian tidak dapat digeneralisasi ke semua disiplin ilmu. Meskipun memiliki batasan tertentu, penelitian ini diharapkan mampu memberikan kontribusi dalam mendukung proses penelusuran referensi ilmiah secara lebih efisien.

1.2 Tujuan Penelitian

Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi dan rekomendasi dokumen karya ilmiah berbasis teks menggunakan model berbasis representasi kontekstual yang telah dioptimalkan melalui *fine-tuning*, dengan penyesuaian *hyperparameter* pada tugas klasifikasi. Evaluasi dilakukan dengan membandingkan performa sistem terhadap beberapa pendekatan lain, baik dalam klasifikasi kategori laboratorium maupun rekomendasi dokumen yang relevan.

1.3 Manfaat Penelitian

Penelitian ini bermanfaat untuk mengevaluasi efektivitas pendekatan klasifikasi dan rekomendasi dokumen ilmiah serta mengidentifikasi metode yang paling sesuai untuk data berbasis teks di bidang ilmu komputer dan informatika. Hasil yang diperoleh dapat dimanfaatkan oleh mahasiswa, dosen, dan peneliti untuk mengoptimalkan akses terhadap literatur ilmiah dengan cara memudahkan pencarian referensi, mempercepat proses klasifikasi dokumen, serta meningkatkan relevansi rekomendasi bacaan.