

BAB I

PENDAHULUAN

Bab pendahuluan ini membahas mengenai latar belakang masalah, rumusan masalah, tujuan, ruang lingkup, serta sistematika penulisan yang akan digunakan dalam dokumen skripsi yang berjudul “Studi Komparatif Representasi Teks TF-IDF dan IndoBERT Berbasis *Content-Based Filtering* Menggunakan *Cosine Similarity* untuk Rekomendasi Destinasi Wisata”.

1.1 Latar Belakang

Sistem rekomendasi merupakan salah satu komponen penting dalam kehidupan digital modern karena berperan membantu pengguna menemukan informasi yang relevan di tengah banyaknya pilihan yang tersedia. Pada sektor pariwisata di Indonesia, keberagaman destinasi wisata yang tersebar di berbagai daerah menawarkan potensi besar dalam meningkatkan pengalaman wisatawan. Namun, banyaknya alternatif destinasi yang tersedia serta penyajian informasi yang tidak terstruktur sering kali menyulitkan wisatawan dalam menentukan pilihan yang sesuai dengan preferensi pribadi, sehingga dapat menurunkan efektivitas pengambilan keputusan dan tingkat kepuasan wisatawan (Solano dkk., 2024).

Dalam pengembangan sistem rekomendasi, terdapat dua pendekatan utama yang umum digunakan, yaitu *Collaborative Filtering* dan *Content-Based Filtering*. *Collaborative Filtering* menghasilkan rekomendasi berdasarkan pola kesamaan preferensi antar pengguna, sedangkan *Content-Based Filtering* menghasilkan rekomendasi dengan menganalisis kemiripan antar item berdasarkan atribut atau konten deskriptif yang dimilikinya (Vall dkk., 2019). Pendekatan *Content-Based Filtering* efektif diterapkan pada domain informasi deskriptif, termasuk pariwisata, karena mampu memanfaatkan konten teks sebagai sumber utama rekomendasi (Mustaqim dkk., 2022). Pada pendekatan *Collaborative Filtering* memiliki keterbatasan dari data interaksi pengguna. Kondisi ini dikenal sebagai *cold start problem*, yaitu keadaan ketika sistem belum memiliki cukup informasi untuk menghasilkan rekomendasi yang akurat bagi pengguna baru. Sebaliknya, *Content-Based Filtering* relatif lebih tahan terhadap permasalahan tersebut karena rekomendasi dihasilkan berdasarkan karakteristik konten item, bukan ketergantungan pada data pengguna lain, sehingga lebih sesuai diterapkan pada kondisi data pengguna yang terbatas (Zhang dkk., 2019).

Representasi teks berperan penting dalam mentransformasikan informasi tekstual menjadi vektor numerik yang dapat diolah secara komputasional. Studi yang dilakukan oleh Mustaqim dkk. (2022) menunjukkan bahwa pemilihan metode representasi teks yang berbeda dapat memberikan dampak signifikan terhadap performa sistem rekomendasi berbasis teks, terutama pada domain yang mengandalkan deskripsi konten sebagai sumber utama informasi. Berbagai metode representasi teks telah dikembangkan, mulai dari pendekatan berbasis statistik hingga pendekatan berbasis *deep learning*. Metode statistik seperti TF-IDF masih banyak digunakan sebagai *baseline* karena efisiensi dan kemudahannya dalam implementasi, sementara model bahasa berbasis Transformer seperti BERT mampu menangkap makna kontekstual dan hubungan semantik antar kata secara lebih mendalam. Namun, penelitian sebelumnya menunjukkan bahwa tidak terdapat satu metode representasi teks yang unggul secara mutlak pada seluruh domain dan karakteristik data, sehingga studi komparatif diperlukan untuk memahami kelebihan dan keterbatasan masing-masing pendekatan (Aggarwal, 2016; Zhang dkk., 2019; Koto dkk., 2020). Berdasarkan pertimbangan tersebut, penelitian ini dibatasi pada kajian komparatif pendekatan representasi teks yang memiliki karakteristik berbeda, yaitu TF-IDF dan IndoBERT. TF-IDF merepresentasikan pendekatan statistik yang umum digunakan sebagai *baseline* dalam sistem rekomendasi berbasis teks, sedangkan IndoBERT merepresentasikan pendekatan *deep learning* kontekstual yang dikembangkan khusus untuk bahasa Indonesia. Pembatasan ini dilakukan secara sadar agar analisis dapat dilakukan secara lebih mendalam dan terkontrol, tanpa bertujuan mengklaim metode terbaik secara universal.

Setelah representasi teks diperoleh, diperlukan metode pengukuran kemiripan antar vektor untuk menentukan tingkat kesamaan konten antar destinasi wisata. Pada representasi teks berbasis TF-IDF, vektor yang dihasilkan bersifat *sparse*, yaitu vektor berdimensi tinggi dengan banyak nilai nol karena hanya sebagian kecil kata yang muncul pada setiap dokumen, sehingga pengukuran jarak menjadi kurang stabil pada ruang vektor berdimensi tinggi (Aggarwal, 2016; Zhang dkk., 2019). Oleh karena itu, penelitian ini menggunakan *Cosine Similarity* yang mengukur kesamaan berdasarkan sudut antar vektor, bukan berdasarkan jarak absolutnya. Pendekatan dengan *Cosine Similarity* lebih *robust* terhadap perbedaan panjang vektor dan sesuai digunakan pada representasi teks berdimensi tinggi, baik yang bersifat *sparse* seperti TF-IDF maupun representasi dense seperti *embedding* IndoBERT. Dengan demikian, *Cosine Similarity* memberikan dasar pengukuran yang konsisten dalam membandingkan

kemiripan konten pada kedua pendekatan representasi teks tersebut (Aggarwal, 2016; Zhang dkk., 2019).

Evaluasi performa sistem rekomendasi dilakukan menggunakan metrik *Normalized Discounted Cumulative Gain* (NDCG), yang mempertimbangkan urutan rekomendasi serta tingkat relevansi berdasarkan *rating* pengguna. Metrik NDCG banyak digunakan dalam penelitian sistem rekomendasi karena mampu merefleksikan kualitas peringkat rekomendasi secara lebih realistis. Melalui evaluasi menggunakan NDCG, penelitian ini bertujuan menganalisis pengaruh perbedaan representasi teks terhadap kualitas rekomendasi destinasi wisata secara sistematis dan terukur (Järvelin & Kekäläinen, 2002).

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan di atas maka rumusan masalah yang akan dibahas oleh penulis yaitu:

1. Bagaimana proses penerapan metode menggunakan representasi teks TF-IDF dan IndoBERT dalam menghasilkan rekomendasi destinasi wisata berdasarkan deskripsi dan kategori konten?
2. Bagaimana pengaruh variasi parameter n-gram pada TF-IDF terhadap kualitas rekomendasi destinasi wisata yang dihasilkan?
3. Bagaimana pengaruh penggunaan IndoBERT dalam kondisi *pre-trained* dibandingkan dengan IndoBERT yang telah melalui proses *Fine-Tuning* terhadap kualitas rekomendasi destinasi wisata?
4. Bagaimana pengaruh variasi nilai *learning rate* pada proses *Fine-Tuning* IndoBERT terhadap performa sistem rekomendasi destinasi wisata?
5. Metode representasi teks manakah yang memberikan performa terbaik dalam mendukung sistem rekomendasi destinasi wisata berdasarkan nilai evaluasi NDCG?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah membangun sistem rekomendasi destinasi wisata menggunakan metode *Content-Based Filtering* dengan membandingkan pendekatan representasi teks TF-IDF, IndoBERT *pretrained*, dan IndoBERT dengan *Fine-Tuning*. Teknik representasi teks ini digunakan untuk menghasilkan representasi numerik dari deskripsi dan kategori objek wisata sebelum dihitung tingkat kemiripannya menggunakan *Cosine Similarity*.

Penelitian ini juga bertujuan mengevaluasi kinerja sistem rekomendasi yang dibangun dengan kedua pendekatan tersebut menggunakan metrik *Normalized Discounted Cumulative Gain* (NDCG), sehingga dapat diketahui metode representasi teks mana yang memberikan rekomendasi yang lebih relevan dan akurat.

1.4 Ruang Lingkup

Penelitian ini memerlukan batasan-batasan supaya penelitian lebih terarah. Batasan-batasannya sebagai berikut:

1. Dataset penelitian ini diperoleh dari dataset publik yang tersedia pada platform Kaggle, yaitu *Indonesian Tourism Destination* (Prabowo, 2021). Dataset terdiri dari 437 destinasi wisata yang direpresentasikan melalui data konten, yaitu deskripsi dan kategori destinasi. Selain data konten, dataset ini juga terdapat 10.000 entri penilaian destinasi wisata dari pengguna serta 300 entri data pengguna yang dimanfaatkan sebagai data dasar (*baseline*) dalam proses evaluasi kinerja sistem rekomendasi menggunakan metrik *Normalized Discounted Cumulative Gain* (NDCG).
2. Pengujian dilakukan dengan representasi teks TF-IDF, IndoBERT *pretrained*, dan IndoBERT dengan *Fine-Tuning* untuk membandingkan kualitas representasi dalam menghasilkan rekomendasi.
3. *TF-IDFVectorizer* dari *scikit-learn* digunakan untuk menghasilkan representasi berbasis frekuensi kata. Pada tahap representasi teks menggunakan TF-IDF Vectorizer, dilakukan pengujian terhadap beberapa variasi parameter n-gram, yaitu unigram (1,1), kombinasi unigram–bigram (1,2), kombinasi unigram–bigram–trigram (1,3), serta bigram (2,2).
4. IndoBERT berbasis arsitektur Transformer dalam kondisi *pre-trained* digunakan sebagai pendekatan representasi teks untuk menghasilkan representasi kontekstual.
5. IndoBERT dengan proses *Fine-Tuning* digunakan sebagai pendekatan representasi teks kontekstual yang disesuaikan dengan karakteristik dataset penelitian.
6. Pada IndoBERT dengan *Fine-Tuning* dilakukan beberapa pengujian *learning rate*, yaitu 2×10^{-5} , 3×10^{-5} , 5×10^{-5} , dan 1×10^{-4} .
7. Pembuktian komputasi matematis dilakukan pada salah satu tempat yaitu tempat dengan *place id* 392 (Ekowisata Mangrove Wonorejo).

8. Sistem rekomendasi yang dibangun dengan pendekatan *Content-Based Filtering*, dengan penerapan *Cosine Similarity* antar destinasi wisata, tanpa memanfaatkan interaksi eksplisit antar pengguna.
9. Evaluasi sistem rekomendasi dilakukan menggunakan metrik *Normalized Discounted Cumulative Gain* (NDCG) untuk menilai kualitas peringkat rekomendasi berdasarkan relevansi *rating* pengguna.

1.5 Sistematika Penulisan

Sistematika penulisan dibagi menjadi beberapa pokok pembahasan. Adapun sistematika penulisan pada laporan skripsi ini sebagai berikut:

BAB I. PENDAHULUAN

Bab ini berisi uraian mengenai latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, ruang lingkup penelitian, serta sistematika penulisan skripsi berjudul “Studi Komparatif Representasi Teks TF-IDF dan IndoBERT Berbasis *Content-Based Filtering* Menggunakan *Cosine Similarity* untuk Rekomendasi Destinasi Wisata”.

BAB II. LANDASAN TEORI

Bab ini membahas landasan teori-teori yang mendasari penelitian, meliputi pembahasan mengenai representasi teks seperti metode berbasis frekuensi TF-IDF dan representasi teks berbasis Transformer melalui IndoBERT *pretrained*, dan IndoBERT dengan proses *Fine-Tuning* digunakan sebagai pendekatan representasi teks kontekstual yang disesuaikan dengan karakteristik dataset penelitian. Selanjutnya, dibahas konsep dasar sistem rekomendasi beserta kategori metode rekomendasi, dengan pendekatan *Content-Based Filtering* menggunakan pengukuran tingkat kemiripan antar item dengan *Cosine Similarity* yang memanfaatkan representasi teks dalam menghasilkan rekomendasi. Bab ini juga menjelaskan metrik evaluasi *Normalized Discounted Cumulative Gain* (NDCG) sebagai indikator utama dalam menilai kualitas hasil rekomendasi.

BAB III. METODOLOGI PENELITIAN

Bab ini menguraikan metodologi penelitian yang digunakan, mencakup jenis penelitian, objek penelitian berupa destinasi wisata di beberapa kota di Indonesia, serta sumber dan karakteristik data yang digunakan. Tahapan pengolahan data dijelaskan mulai dari *text preprocessing* dengan representasi teks menggunakan TF-IDF dan IndoBERT *pretrained*, dan IndoBERT dengan *Fine-Tuning*, perhitungan kemiripan menggunakan *Cosine Similarity*. Selain itu, prosedur evaluasi sistem menggunakan metrik NDCG dijelaskan secara sistematis untuk menilai performa model dalam menghasilkan rekomendasi.

BAB IV. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil dan pembahasan dari eksperimen yang telah dilakukan. Pembahasan meliputi analisis tahap *preprocessing* data, hasil perbandingan representasi teks menggunakan TF-IDF, IndoBERT *pre-trained*, dan IndoBERT dengan proses *Fine-Tuning*. Selain itu, dipaparkan juga hasil perhitungan *Cosine Similarity* serta rekomendasi destinasi wisata yang dihasilkan oleh sistem. Evaluasi kinerja sistem rekomendasi dilakukan menggunakan metrik *Normalized Discounted Cumulative Gain* (NDCG), yang diikuti dengan interpretasi hasil dan analisis komparatif antar metode representasi teks yang diuji.

BAB V. KESIMPULAN

Bab ini berisi kesimpulan dari penelitian yang telah dilakukan, yang merangkum hasil analisis dan evaluasi serta saran untuk pengembangan penelitian lebih lanjut.