

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi digital telah memberikan pengaruh signifikan terhadap perubahan perilaku pengguna dalam memperoleh informasi, termasuk dalam proses pencarian dan pemilihan produk melalui berbagai platform digital. Perubahan tersebut juga terjadi pada sektor *fitness* dan olahraga yang mengalami peningkatan variasi produk, mulai dari peralatan latihan, perlengkapan olahraga, hingga produk pendukung aktivitas kebugaran. Peningkatan jumlah dan variasi produk memberikan alternatif pilihan yang lebih luas bagi pengguna, tetapi pada sisi lain dapat meningkatkan kompleksitas dalam menentukan produk yang sesuai dengan kebutuhan dan preferensi pengguna. Kanwal dkk. (2021) menjelaskan bahwa peningkatan jumlah informasi berbasis teks pada platform digital dapat menyebabkan *information overload*, sehingga diperlukan suatu mekanisme penyaringan informasi yang mampu menghasilkan rekomendasi secara efektif dan relevan.

Sistem rekomendasi merupakan salah satu pendekatan yang dapat digunakan untuk mengatasi permasalahan tersebut melalui proses penyaringan dan pemeringkatan informasi berdasarkan karakteristik tertentu. Saifudin & Widiyaningtyas. (2024) menyatakan bahwa sistem rekomendasi bertujuan membantu pengguna memperoleh informasi yang relevan dengan melakukan analisis terhadap karakteristik suatu *item*. Salah satu pendekatan yang banyak digunakan dalam pengembangan sistem rekomendasi adalah *content-based filtering*. Khadka & Sthapit. (2025) menjelaskan bahwa pendekatan *content-based filtering* menghasilkan rekomendasi berdasarkan tingkat kemiripan karakteristik suatu *item* dengan preferensi pengguna. Pendekatan ini memiliki keunggulan karena tidak bergantung pada riwayat interaksi pengguna lain, sehingga sesuai diterapkan pada kondisi ketika informasi interaksi pengguna belum tersedia secara optimal. Pada domain produk *fitness* dan olahraga, pendekatan ini dapat

diterapkan dengan memanfaatkan informasi tekstual produk, seperti nama produk, kategori, dan ulasan pengguna sebagai representasi karakteristik produk.

Meskipun pendekatan *content-based filtering* telah banyak diterapkan dalam berbagai sistem rekomendasi, terdapat tantangan dalam proses representasi informasi tekstual produk. Sebagian penelitian sebelumnya masih menggunakan pendekatan berbasis frekuensi kata, seperti *bag-of-words* dan TF-IDF, yang memiliki keterbatasan dalam memahami hubungan semantik antar kata. Selain itu, beberapa sistem rekomendasi produk masih memanfaatkan informasi eksplisit berupa *rating*, riwayat transaksi, maupun pola interaksi pengguna. Pendekatan tersebut dapat mengalami keterbatasan ketika data interaksi pengguna tidak tersedia atau jumlah interaksi masih rendah. Oleh karena itu, diperlukan suatu pendekatan yang mampu merepresentasikan informasi tekstual produk dengan mempertahankan hubungan makna antar dokumen sehingga proses pencarian kemiripan produk dapat dilakukan secara lebih efektif.

Pada penelitian ini, pendekatan *content-based filtering* diterapkan menggunakan *dataset* publik *Amazon Reviews'23 Dataset* yang berisi informasi produk dan ulasan pengguna. *Dataset* tersebut memiliki beberapa atribut, yaitu *Product_id*, *Product_name*, *category*, dan *clean_text*. Atribut *Product_id* digunakan sebagai identitas produk, sedangkan atribut *Product_name*, *category*, dan *clean_text* digunakan sebagai representasi konten produk dalam proses rekomendasi. Sebelum digunakan dalam proses pemodelan, data melalui tahapan *preprocessing* yang meliputi *data selection*, *cleaning text*, *case folding*, *tokenizing*, dan *stopword removal*. Tahapan tersebut dilakukan untuk menghasilkan data teks yang lebih terstruktur sehingga dapat digunakan dalam proses pembentukan representasi dokumen.

Tantangan utama dalam pengolahan data berbasis teks adalah menghasilkan representasi numerik yang mampu mempertahankan informasi semantik dari suatu dokumen. Metode representasi teks konvensional seperti *bag-of-words* dan TF-IDF cenderung berfokus pada frekuensi kemunculan kata sehingga belum mampu menangkap hubungan makna antar kata secara mendalam. Moskalenko dkk. (2020) menjelaskan bahwa pendekatan *semantic embedding* seperti *Doc2Vec*

mampu menghasilkan representasi dokumen dalam bentuk vektor dengan mempertimbangkan hubungan semantik pada suatu teks. Selanjutnya, Iskandar & Kurniawati. (2025) menunjukkan bahwa *Doc2Vec* mampu digunakan untuk merepresentasikan dokumen dengan mempertimbangkan kemiripan makna meskipun terdapat perbedaan struktur kata. Berdasarkan kemampuan tersebut, penelitian ini menggunakan *Doc2Vec* untuk menghasilkan representasi vektor dari konten produk yang terdiri atas nama produk, kategori, dan ulasan pengguna.

Setelah proses pembentukan representasi dokumen dilakukan, diperlukan metode pengukuran kemiripan untuk menentukan tingkat kesesuaian antara *query* pengguna dan dokumen produk. Penelitian ini menggunakan *cosine similarity* karena metode tersebut mampu mengukur tingkat kedekatan antar vektor berdasarkan orientasi ruang representasi. Karpukhin dkk. (2020) menjelaskan bahwa pendekatan berbasis *dense embedding* yang dikombinasikan dengan metode pengukuran kemiripan vektor dapat digunakan untuk mengidentifikasi hubungan semantik antar dokumen. Nilai kemiripan yang diperoleh selanjutnya digunakan sebagai dasar dalam menentukan peringkat rekomendasi produk berdasarkan tingkat kesesuaian konten terhadap *query* pengguna.

Berdasarkan permasalahan tersebut, penelitian ini mengembangkan sistem rekomendasi produk *fitness* dan olahraga berbasis *content-based filtering* menggunakan metode *Doc2Vec* dan *cosine similarity*. Kontribusi penelitian ini terletak pada penerapan pendekatan berbasis representasi semantik untuk mengolah informasi tekstual produk, yaitu nama produk, kategori, dan ulasan pengguna, sebagai dasar dalam menghasilkan rekomendasi. Pendekatan tersebut memungkinkan model menghasilkan representasi produk berdasarkan hubungan semantik dari informasi tekstual, kemudian dikombinasikan dengan mekanisme pencarian dan validasi pada sisi aplikasi untuk menghasilkan rekomendasi yang sesuai. Kinerja sistem yang dikembangkan dievaluasi menggunakan metrik *Precision*, *Recall*, dan *Normalized Discounted Cumulative Gain* (NDCG) untuk mengukur kemampuan sistem dalam menghasilkan rekomendasi produk yang relevan.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, dapat diketahui bahwa permasalahan yang dirumuskan dalam penelitian ini adalah sebagai berikut:

1. Bagaimana membangun sistem rekomendasi produk *fitness* dan olahraga berbasis *content-based filtering* dengan memanfaatkan informasi nama produk, kategori, dan ulasan pengguna melalui tahapan *preprocessing*, representasi teks menggunakan *Doc2Vec*, serta pengukuran kemiripan menggunakan *cosine similarity*?
2. Bagaimana kinerja sistem rekomendasi produk *fitness* dan olahraga berbasis *content-based filtering* menggunakan *Doc2Vec* dan *cosine similarity* berdasarkan hasil evaluasi rekomendasi dengan metrik *Precision*, *Recall*, dan *NDCG*?

1.3 Tujuan Penelitian

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

1. Membangun sistem rekomendasi produk *fitness* dan olahraga berbasis *content-based filtering* dengan memanfaatkan informasi nama produk, kategori, dan ulasan pengguna melalui tahapan *preprocessing*, representasi teks menggunakan *Doc2Vec*, serta pengukuran tingkat kemiripan menggunakan *cosine similarity*.
2. Mengetahui kinerja sistem rekomendasi produk *fitness* dan olahraga berbasis *content-based filtering* menggunakan *Doc2Vec* dan *cosine similarity* melalui evaluasi menggunakan metrik *Precision*, *Recall*, dan *NDCG*.

1.4 Manfaat Penelitian

1. Memberikan kemudahan bagi pengguna dalam menemukan produk *fitness* dan olahraga yang sesuai berdasarkan kebutuhan melalui sistem rekomendasi berbasis *content-based filtering*.
2. Membantu pengguna mengurangi waktu pencarian produk dengan menyediakan rekomendasi berdasarkan kesamaan informasi nama produk, kategori, dan ulasan pengguna.

3. Menghasilkan sistem rekomendasi produk *fitness* dan olahraga berbasis *website* yang mampu mengolah data teks menggunakan metode *Doc2Vec* dan *cosine similarity*.
4. Memberikan referensi penerapan metode *Doc2Vec* dan *cosine similarity* dalam pengembangan sistem rekomendasi berbasis data tekstual.
5. Menjadi bahan referensi bagi penelitian selanjutnya yang berkaitan dengan sistem rekomendasi, pengolahan teks, dan penerapan metode *semantic embedding*.

1.5 Batasan Penelitian

Batasan masalah pada penelitian ini ditetapkan agar pembahasan lebih terfokus dan sesuai dengan tujuan penelitian. Adapun batasan penelitian ini adalah sebagai berikut:

1. *Dataset* yang digunakan dalam penelitian ini berasal dari *Amazon Reviews*'23 *Dataset* yang dikembangkan oleh McAuley Lab pada tahun 2023. *Dataset* tersebut berupa data ulasan produk (*user Reviews*) dan informasi produk (*item metadata*) yang kemudian diseleksi berdasarkan kategori *fitness* dan olahraga.
2. Data yang digunakan dalam penelitian telah melalui tahap *preprocessing* untuk mempersiapkan data teks sebelum digunakan dalam sistem rekomendasi. Tahapan *preprocessing* meliputi *data selection*, *cleaning text*, *case folding*, *tokenizing*, dan *stopword removal*.
3. Atribut *dataset* yang digunakan dalam proses rekomendasi meliputi *Product_name*, *category*, dan *clean_text*, sedangkan atribut *Product_id* digunakan sebagai identitas produk.
4. Informasi produk yang digunakan sebagai dasar rekomendasi hanya berasal dari karakteristik konten produk berupa nama produk, kategori, dan ulasan pengguna, tanpa menggunakan data interaksi pengguna seperti riwayat pembelian, jumlah klik, maupun preferensi pengguna lain.
5. Sistem rekomendasi yang dikembangkan menggunakan pendekatan *content-based filtering*, yaitu memberikan rekomendasi berdasarkan tingkat kemiripan konten antara *query* pengguna dengan informasi produk yang tersedia.

6. Metode yang digunakan untuk menghasilkan representasi teks adalah *Doc2Vec*, yang digunakan untuk mengubah informasi tekstual produk dan *query* pengguna menjadi bentuk vektor numerik.
7. Pengukuran tingkat kemiripan antara vektor *query* pengguna dan vektor dokumen produk dilakukan menggunakan metode *cosine similarity*.
8. Hasil rekomendasi yang diberikan sistem berupa daftar produk yang diurutkan berdasarkan nilai kemiripan (*similarity score*) tertinggi.
9. Implementasi sistem dilakukan dalam bentuk *website* yang digunakan sebagai media pengguna dalam memasukkan *query* dan melihat hasil rekomendasi produk.
10. Evaluasi kinerja sistem dilakukan menggunakan metrik *Precision*, *Recall*, dan *Normalized Discounted Cumulative Gain* (NDCG) untuk mengukur kualitas hasil rekomendasi yang dihasilkan.
11. Penelitian ini tidak membahas proses transaksi, pembayaran, pengelolaan stok, maupun integrasi langsung dengan platform *e-commerce*.