BAB 1
PENDAHULUAN

1.1 Latar Belakang

Perkembangan industri film dan layanan streaming digital mengalami peningkatan yang signifikan dalam beberapa tahun terakhir. Berbagai penelitian menunjukkan bahwa ribuan judul film baru dirilis setiap tahun dan platform streaming besar menyediakan katalog berisi puluhan ribu konten film dan serial yang dapat diakses oleh pengguna secara daring. Kondisi tersebut memperluas pilihan tontonan, tetapi di sisi lain juga memunculkan fenomena kelebihan informasi (*information overload*) yang menyulitkan pengguna dalam menentukan film yang sesuai dengan minat personal (Airen & Agrawal, 2021; Fajriansyah dkk., 2021). Dalam konteks ini, sistem temu kembali informasi (*Information Retrieval*) berperan penting sebagai bagian dari sistem informasi untuk membantu menyaring dan menemukan konten secara spesifik berdasarkan kueri pengguna. Hal tersebut menjadikan sistem temu kembali informasi tidak lagi sekadar fitur tambahan, melainkan menjadi komponen strategis dalam pengelolaan informasi pada platform digital (Yao, 2023).

Meskipun sistem pencarian telah banyak diterapkan, pengguna masih sering dihadapkan pada fenomena *filter bubble*, di mana sistem cenderung hanya mengurung pengguna pada suguhan film-film yang sedang populer atau terpaku pada genre dasar yang sama secara berulang. Penentuan film yang sekadar berdasarkan tren popularitas sering menghasilkan daftar pencarian yang bersifat umum dan belum tentu selaras dengan preferensi spesifik alur cerita masing-masing pengguna. Akibatnya, pengguna perlu meluangkan waktu lebih lama untuk menelusuri berbagai pilihan sebelum menemukan film yang dianggap paling sesuai. Kondisi tersebut dikenal sebagai *choice paralysis*, yaitu situasi ketika individu mengalami kesulitan dalam mengambil keputusan akibat terlalu banyak pilihan umum yang tersedia (Romero Meza & D'Urso, 2024). Oleh karena itu, sistem yang dibangun dalam penelitian ini dirancang untuk memberikan nilai tambah (*value*) bagi pengguna melalui sistem temu kembali informasi murni berbasis kata kunci (*keyword*). Dengan mengetikkan kueri kata kunci secara eksplisit yang

mendeskripsikan jalan cerita yang diinginkan, sistem dapat membantu pengguna menemukan *hidden gems* atau film-film yang secara semantik benar-benar relevan dengan kedalaman alur cerita yang disukai. Hal ini menunjukkan bahwa pemahaman mendalam terhadap preferensi spesifik pengguna masih menjadi tantangan dalam praktik sistem temu kembali informasi film saat ini (Velamentosa & Zuliarso, 2025).

Selain pemahaman terhadap preferensi pengguna, tantangan utama dalam sistem temu kembali informasi berbasis *content-based filtering* adalah ketersediaan dan kekayaan teks deskripsi film. Penggunaan dataset publik sejatinya telah memberikan kontribusi yang sangat besar dalam penelitian karena memiliki struktur yang terstandarisasi, keandalan yang tinggi, dan memuat informasi dasar yang sangat baik untuk melatih model algoritma (Sinha & Sharma, 2023). Kelengkapan teks ini menjadi faktor penting untuk meminimalisasi kendala *cold start* pada film-film baru, di mana film yang belum memiliki riwayat interaksi penonton tetap dapat ditemukan secara relevan asalkan memiliki deskripsi konten yang memadai (Natarajan dkk., 2020). Namun, untuk merepresentasikan makna semantik secara maksimal, model bahasa membutuhkan volume data teks (korpus) yang masif dan kaya akan variasi kata. Untuk semakin mengoptimalkan potensi model tersebut, dataset publik yang unggul ini perlu didukung dengan tambahan teks yang lebih mendetail dan mutakhir. Kondisi tersebut memicu kebutuhan akan integrasi mekanisme ekstraksi data secara mandiri, seperti *web scraping* langsung dari IMDb, sebagai langkah komplementer. Pemanfaatan *web scraping* diakui sebagai pendekatan metodologis yang sah untuk memperluas cakupan dataset teks, sehingga sistem memiliki basis data yang lebih kaya untuk memahami konteks cerita film (Diouf dkk., 2019).

Pendekatan yang umumnya diterapkan untuk mencari dan menemukan item berdasarkan karakteristik fitur atau deskripsinya adalah *content-based filtering*. Dalam konteks temu kembali informasi, pendekatan ini dimodifikasi untuk menyeleksi dan mengurutkan film berdasarkan tingkat kemiripan antara kueri kata kunci yang dimasukkan pengguna secara eksplisit dengan deskripsi film yang tersedia (Fajriansyah dkk., 2021). Dalam metode ini, ekstraksi teks yang sering digunakan masih berbasis frekuensi kata seperti TF-IDF. Namun, pendekatan

tersebut memiliki keterbatasan dalam memahami konteks makna naratif dari sebuah deskripsi film. Kata-kata yang berbeda secara leksikal tetapi memiliki makna serupa dapat dianggap tidak berhubungan oleh metode TF-IDF, sehingga kemiripan yang dihasilkan belum sepenuhnya mencerminkan kesamaan isi (Nguyen dkk., 2020).

Sebagai alternatif yang lebih efektif untuk mengatasi keterbatasan tersebut, *word embeddings* diperkenalkan untuk merepresentasikan teks ke dalam ruang vektor dengan mempertahankan makna semantis dan hubungan antarkata (Mikolov dkk., 2013). Pendekatan ini memungkinkan sistem untuk memahami konteks naratif secara lebih menyeluruh. Dengan mengintegrasikan *word embeddings*, representasi kata kunci dari pengguna dapat dipetakan secara lebih presisi dengan vektor deskripsi film dibandingkan metode berbasis frekuensi kata. Lebih lanjut, pencarian ini beroperasi murni berdasarkan masukan kata kunci (*keyword*) tanpa bergantung pada judul film acuan (*seed movie*). Secara komputasi, *word embeddings* mengubah teks menjadi representasi vektor angka, di mana vektor kata kunci bertindak sebagai variabel A dan vektor deskripsi film sebagai variabel B. Tingkat kemiripan antara kedua variabel numerik tersebut kemudian dihitung secara langsung menggunakan rumus *Cosine Similarity*. Perhitungan ini akan mengurutkan film berdasarkan nilai relevansi tertinggi untuk menghasilkan daftar pencarian yang sesuai dengan deskripsi pengguna.

Meskipun beberapa penelitian telah mengimplementasikan *content-based filtering* dan *word embeddings* dalam sistem temu kembali informasi, sebagian besar masih berfokus pada pemanfaatan *dataset* publik sekunder secara tunggal dan bergantung pada judul film acuan dalam pencariannya (Nguyen dkk., 2020; Sinha dan Sharma, 2023). Hal ini menyisakan celah untuk mengombinasikan *dataset* publik dengan *web scraping* ke dalam sebuah sistem berbasis web. Pencariannya juga diubah menjadi murni menggunakan *input keyword* bahasa Indonesia yang nantinya akan diterjemahkan secara otomatis oleh sistem. Oleh karena itu, penelitian ini mengintegrasikan *word embeddings* dan *Cosine Similarity* dalam sistem temu kembali informasi tanpa dokumen acuan (*seed movie*). Kualitas urutan dokumen pencarian diukur menggunakan metrik *Top-N Retrieval* (*Precision@K*, *Recall@K*, dan MRR). Sebagai *ground truth*, status relevansi film terhadap

keyword divalidasi melalui penilaian ahli (*expert judgment*). Integrasi ini diharapkan menghasilkan algoritma adaptif yang menempatkan film paling relevan di urutan teratas.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana menerapkan metode *content-based filtering* dengan *word embeddings* pada deskripsi film untuk membangun algoritma sistem temu kembali informasi yang memproses data melalui tahapan *training* dan *testing*, menerima masukan murni berupa kueri kata kunci (*keyword*) pengguna tanpa bergantung pada judul film acuan (*seed movie*), dan menghasilkan daftar pencarian berdasarkan perhitungan *Cosine Similarity*?
2. Bagaimana tingkat relevansi urutan pencarian yang dihasilkan oleh algoritma tersebut ketika dievaluasi menggunakan metrik *Top-N Retrieval* (*Precision@K, Recall@K,* dan *Mean Reciprocal Rank* / MRR) dengan acuan *ground truth* berdasarkan validasi pakar (*expert judgment*)?

1.3 Tujuan Penelitian

Sejalan dengan rumusan masalah yang telah ditetapkan, tujuan dari penelitian ini adalah sebagai berikut:

1. Menerapkan metode *content-based filtering* dengan *word embeddings* pada deskripsi film untuk membangun algoritma sistem temu kembali informasi yang memproses data melalui tahapan *training* dan *testing*, menerima masukan murni berupa kueri kata kunci (*keyword*) pencarian dari pengguna tanpa bergantung pada judul film acuan (*seed movie*), dan menghasilkan daftar urutan pencarian informasi berdasarkan perhitungan *Cosine Similarity*.
2. Mengukur dan menganalisis tingkat relevansi urutan pencarian yang dihasilkan oleh algoritma tersebut menggunakan metrik *Top-N Retrieval* (*Precision@K, Recall@K,* dan *Mean Reciprocal Rank* / MRR) dengan acuan *ground truth* berdasarkan validasi pakar (*expert judgment*).

1.4 Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberikan manfaat, baik secara teoretis maupun praktis, di antaranya sebagai berikut:

1. Manfaat Teoritis

   Penelitian ini diharapkan dapat memberikan kontribusi keilmuan di bidang Sistem Informasi dan Ilmu Komputer, khususnya terkait penerapan dan evaluasi kinerja sistem temu kembali informasi berbasis teks. Temuan penelitian ini dapat menjadi referensi empiris mengenai efektivitas pemanfaatan *word embeddings* dalam merepresentasikan makna semantik deskripsi film dibandingkan dengan metode frekuensi kata konvensional. Selain itu, penelitian ini juga memperkaya literatur terkait penerapan pencarian murni berbasis kata kunci (*keyword*) tanpa judul film acuan (*seed movie*), serta evaluasi tingkat relevansi urutan pencarian yang diukur menggunakan metrik *Top-N Retrieval* (*Precision@K, Recall@K*, dan MRR).

2. Manfaat Praktis

   a. Bagi Pengguna: Penelitian ini diharapkan dapat memberikan kemudahan bagi pengguna dalam menemukan film dengan kemiripan kedalaman alur cerita yang relevan sesuai preferensi personal melalui pencarian berbasis kata kunci (*keyword*) tanpa perlu mengetahui atau memasukkan judul film spesifik. Hal ini bertujuan untuk meminimalisasi kesulitan pengguna dalam memilih tontonan (*choice paralysis*) serta membantu pengguna menemukan *hidden gems* atau alternatif film berdasarkan minat spesifik, sehingga pengguna tidak hanya terkurung pada hasil pencarian tren populer yang umum disajikan oleh platform digital (*filter bubble*).

   b. Bagi Pengembang Sistem: Penelitian ini memberikan wawasan teknis mengenai implementasi *word embeddings* dan perhitungan *Cosine Similarity* dalam membangun model algoritma *content-based filtering*. Selain itu, penelitian ini menunjukkan bagaimana integrasi antara *dataset* publik dan mekanisme ekstraksi data mandiri (*web scraping*) dapat dimanfaatkan sebagai alternatif penyediaan pangkalan data (korpus) deskripsi film yang lebih kaya dan mutakhir.

1.5 Batasan Masalah

Agar penelitian ini lebih terfokus dan selaras dengan tujuan yang telah ditetapkan, maka batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Objek penelitian dibatasi hanya pada film (*movies*), tidak mencakup serial televisi (*TV series*) atau acara hiburan lainnya.
2. Sistem temu kembali informasi dibangun berbasis web yang digunakan hanya sebagai media visualisasi, menggunakan pendekatan *content-based filtering* dengan algoritma *word embeddings* (Word2Vec) untuk representasi teks dan *Cosine Similarity* untuk menghitung nilai kemiripan antarfilm.
3. Data bersumber dari *dataset* publik sebagai basis utama yang dilengkapi dengan hasil *web scraping* IMDb sebagai data komplementer. Data dibatasi maksimal 5.000 film dengan kriteria *rating* minimal 6.0 yang dirilis hingga April 2026.
4. Pengujian dan evaluasi kinerja sistem dilakukan menggunakan metode pembagian data *Holdout* (80% data untuk *training* dan 20% data untuk *testing*), di mana sistem akan mengambil nilai hasil model pembelajaran yang paling baik (optimal).
5. Evaluasi kinerja sistem difokuskan pada tingkat relevansi daftar urutan keluaran dengan menggunakan metrik *Top-N Retrieval* yaitu *Precision@K*, *Recall@K*, dan *Mean Reciprocal Rank* (MRR).
6. Analisis kemiripan semantik difokuskan murni pada komputasi teks deskripsi/sinopsis film untuk menjaga objektivitas alur cerita, sehingga tidak melibatkan sentimen opini atau ulasan pengguna (*user review*).
7. Pencarian dilakukan murni berdasarkan masukan berupa kata kunci (*keyword*) dari pengguna, tanpa menggunakan input judul film acuan (*seed movie*) maupun data riwayat personal pengguna (*user profiling*).
8. Sistem tidak memfilter atau mengklasifikasikan hasil pencarian berdasarkan batasan usia penonton (*age rating*). Sistem juga tidak merancang algoritma khusus (seperti diversifikasi otomatis) untuk mengatasi fenomena *filter bubble* secara komputasional, melainkan mengandalkan dinamika dan variasi masukan *keyword* dari pengguna untuk memutus rantai pencarian yang monoton.

9. Penelitian ini berfokus murni pada evaluasi performa algoritma temu kembali informasi sehingga tidak mencakup perancangan arsitektur sistem (UML) maupun pengujian fungsionalitas perangkat lunak (*Blackbox Testing*).
10. Pemrosesan bahasa alami (NLP) dieksekusi menggunakan teks berbahasa Inggris. Bahasa Indonesia hanya digunakan pada antarmuka web (*User Interface*) dan evaluasi pakar.
11. Nilai kebenaran acuan (*ground truth*) dievaluasi secara manual oleh 2 (dua) pakar berdasarkan keselarasan konteks alur cerita terhadap kata kunci, yang disepakati melalui mekanisme *Inter-Annotator Agreement*. Evaluasi tingkat relevansi sistem menggunakan *ground truth* pakar dibatasi pada 3 peringkat pencarian teratas (Top-3 Retrieval) demi menjaga efisiensi dan akurasi penilaian akibat keterbatasan kognitif evaluator, sedangkan visualisasi pada antarmuka aplikasi tetap menyajikan 10 peringkat teratas (Top-10).
12. Sistem tidak melakukan tahapan *pre-processing* (pemrosesan awal) terhadap kueri kata kunci (*keyword*) yang diinputkan oleh pengguna. Input diproses secara mentah (*raw data*), sehingga sistem memiliki kerentanan atau berpotensi gagal menemukan hasil yang relevan apabila terdapat kesalahan pengetikan (*typo*), penggunaan singkatan, atau struktur kata yang tidak baku pada kueri pencarian.