

SISTEM REKOMENDASI VIDEO GAME BERBASIS CONTENT BASED FILTERING MENGGUNAKAN TF-IDF DAN COSINE SIMILARITY

Muhammad Farrel Yafiwaghani Firjatullah, Jati Sasongko Wibowo

Teknik Informatika, Universitas Stikubank Semarang
Jl. Kendeng, No. 5, Bendanduwur, Gajah Mungkur, Bendan Duwur,
Kec. Gajahmungkur, Kota Semarang, Jawa Tengah 50233
muhammadfarrellyf@gmail.com

ABSTRAK

Industri video game tumbuh pesat dengan lebih dari 14.000 judul baru dirilis setiap tahun pada platform distribusi digital, sehingga pengguna semakin sulit menemukan game yang sesuai preferensi dari ribuan pilihan yang tersedia. Kondisi ini menciptakan fenomena *information overload* yang tidak dapat diselesaikan secara manual oleh pengguna. Sistem rekomendasi berbasis *Collaborative Filtering* juga tidak mampu bekerja optimal karena keterbatasan data interaksi pengguna (*cold-start problem*). Penelitian ini bertujuan membangun sistem rekomendasi video game berbasis *Content-Based Filtering* yang mampu menghasilkan rekomendasi relevan tanpa bergantung pada data historis pengguna, sekaligus menganalisis pengaruh mekanisme *sorting rating* terhadap kualitas rekomendasi. Sistem dibangun menggunakan algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity* dengan dataset 60.000 judul game dari platform *Backloggd*, direduksi menjadi 37.669 data bersih setelah *preprocessing*. Atribut *Summary*, *Genres*, dan *Platforms* digabungkan sebagai representasi konten dan dikonversi menjadi matriks TF-IDF berukuran 37.669×71.951 . Evaluasi menggunakan *Precision@5* pada empat skenario lintas genre menghasilkan *Mean Precision@5* sebesar 0,95. Mekanisme *sorting Rating* terbukti meningkatkan rata-rata *Rating* rekomendasi dari 3,22 menjadi 3,90 dengan *trade-off* penurunan *Cosine Similarity* rata-rata sebesar 0,0532.

Kata kunci : *Content-Based Filtering, Cosine Similarity, Precision@K, Sistem Rekomendasi, TF-IDF, Video Game*

1. PENDAHULUAN

Perkembangan industri video game dalam beberapa tahun terakhir menunjukkan pertumbuhan yang sangat pesat seiring dengan meningkatnya penetrasi internet dan kehadiran berbagai platform distribusi digital seperti Steam, Epic Games Store, dan PlayStation Network. Pada tahun 2023, lebih dari 14.000 judul game baru dirilis di Steam saja, menjadikan total katalog yang tersedia mencapai puluhan ribu judul. Kondisi ini secara langsung menciptakan fenomena *information overload*, yaitu keadaan ketika jumlah informasi atau pilihan game yang tersedia jauh melampaui kemampuan pengguna untuk mengevaluasinya secara manual. Fenomena ini berdampak pada kesulitan pengguna dalam menemukan game yang sesuai dengan preferensi mereka, terutama ketika ribuan judul dengan genre dan deskripsi beragam tersedia dalam satu platform digital [1].

Sistem rekomendasi hadir sebagai salah satu solusi teknologi untuk mengatasi permasalahan tersebut dengan cara membantu pengguna menemukan item yang relevan berdasarkan preferensi atau karakteristik konten tertentu. Secara umum, terdapat tiga pendekatan utama dalam sistem rekomendasi, yaitu *Collaborative Filtering* yang memanfaatkan data perilaku pengguna lain, *Content-Based Filtering* yang memanfaatkan atribut konten item, dan pendekatan hibrida yang menggabungkan keduanya. Di antara ketiga pendekatan tersebut, *Content-Based Filtering* (CBF) memiliki keunggulan utama karena tidak

bergantung pada data perilaku pengguna lain sehingga lebih stabil digunakan pada kondisi *cold-start*, yakni ketika data interaksi pengguna masih terbatas atau belum tersedia [2].

Dalam implementasi *Content-Based Filtering* berbasis teks, proses pengolahan dan representasi teks menjadi aspek yang sangat menentukan kualitas hasil rekomendasi. Salah satu metode yang paling banyak digunakan untuk merepresentasikan dokumen teks secara numerik adalah *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode ini memberikan bobot pada setiap kata berdasarkan kombinasi frekuensi kemunculannya dalam suatu dokumen dan kejarangannya di seluruh korpus, sehingga kata-kata yang paling diskriminatif dan relevan mendapatkan bobot yang lebih tinggi. Selanjutnya, tingkat kemiripan antar dokumen yang telah direpresentasikan sebagai vektor TF-IDF dihitung menggunakan *Cosine Similarity*, yang mengukur kedekatan arah antara dua vektor dalam ruang multidimensi. Kombinasi TF-IDF dan *Cosine Similarity* telah banyak terbukti efektif dalam berbagai aplikasi sistem rekomendasi berbasis teks [3].

Penelitian terdahulu menunjukkan bahwa penerapan CBF dengan TF-IDF dan *Cosine Similarity* telah berhasil diimplementasikan pada berbagai domain, antara lain sistem rekomendasi film [4], sistem rekomendasi destinasi wisata [5], sistem rekomendasi event [6], dan sistem rekomendasi lagu [7].

Namun demikian, sebagian besar penelitian tersebut masih memiliki keterbatasan pada skala data yang digunakan serta belum mengoptimalkan kombinasi fitur metadata multi-label yang khas pada domain video game, seperti genre multi-kategori dan informasi platform. Selain itu, aspek kualitas penilaian pengguna (*rating*) sebagai filter sekunder pada hasil rekomendasi juga belum banyak dikaji secara sistematis.

Berdasarkan kondisi tersebut, penelitian ini bertujuan mengisi kesenjangan yang ada dengan mengimplementasikan sistem rekomendasi video game berbasis *Content-Based Filtering* menggunakan TF-IDF dan *Cosine Similarity* pada dataset berskala besar yang terdiri dari 60.000 judul game dari platform *Backloggd*. Penelitian ini secara khusus menggabungkan tiga fitur tekstual sekaligus, yaitu *Summary*, *Genres*, dan *Platforms*, serta mengintegrasikan mekanisme sorting berbasis *Rating* sebagai filter kualitas sekunder yang bertujuan memastikan rekomendasi yang dihasilkan tidak hanya relevan secara konten, tetapi juga berkualitas tinggi menurut penilaian komunitas pengguna game [8].

Penggunaan dataset berskala besar juga mengacu pada pendekatan sebelumnya [9] yang menunjukkan pentingnya skala dan kelengkapan atribut dataset dalam meningkatkan kualitas rekomendasi game. Selain itu, penelitian ini mempertimbangkan temuan mengenai pentingnya normalisasi vektor TF-IDF dalam komputasi *Cosine Similarity* [10].

Penelitian ini juga menerapkan rekomendasi bahwa representasi konten yang kaya secara tekstual merupakan fondasi utama keberhasilan sistem rekomendasi berbasis konten [11], [12].

Untuk mengatasi permasalahan tersebut, penelitian ini merancang pipeline sistem rekomendasi empat tahap. Tahap pertama adalah *data preprocessing* untuk membersihkan data dan menstandarisasi dataset dari 60.000 menjadi 37.669 data bersih. Tahap kedua adalah ekstraksi fitur menggunakan TF-IDF Vectorizer yang menggabungkan atribut *Summary*, *Genres*, dan *Platforms* menjadi matriks representasi konten. Tahap ketiga adalah perhitungan matriks *Cosine Similarity* untuk seluruh pasangan game. Tahap keempat adalah mekanisme *sorting* dua-tahap: sistem mengambil top-20 kandidat berdasarkan *Cosine Similarity* tertinggi, kemudian mengurutkan ulang berdasarkan *Rating* tertinggi untuk menghasilkan 5 rekomendasi final yang relevan secara konten sekaligus berkualitas tinggi menurut penilaian komunitas [8], [9], [11].

2. TINJAUAN PUSTAKA

2.1. Sistem Rekomendasi

Sistem rekomendasi merupakan sistem cerdas yang dirancang untuk memberikan saran atau rekomendasi item kepada pengguna berdasarkan preferensi, kebutuhan, atau perilaku pengguna sebelumnya. Sistem ini memainkan peran penting dalam mengatasi permasalahan *information overload*

yang terjadi akibat semakin banyaknya pilihan yang tersedia pada platform digital. Dengan memanfaatkan teknik *data mining* dan *machine learning*, sistem rekomendasi mampu menganalisis pola dan hubungan dalam data untuk menghasilkan rekomendasi yang relevan dan personal. Implementasi sistem rekomendasi banyak diterapkan pada berbagai bidang seperti *e-commerce*, *media streaming*, perpustakaan digital, dan permainan digital. Keberadaan sistem ini terbukti membantu pengguna mengambil keputusan secara lebih efektif dan efisien di tengah melimpahnya pilihan informasi [1].

2.2. Content-Based Filtering

Content-Based Filtering (CBF) merupakan salah satu pendekatan utama dalam sistem rekomendasi yang memberikan rekomendasi berdasarkan kesamaan atribut atau konten dari item yang diminati pengguna. Berbeda dengan *Collaborative Filtering* yang memerlukan data perilaku banyak pengguna, CBF sepenuhnya mengandalkan representasi konten item itu sendiri, seperti deskripsi teks, kategori, tag, atau metadata lainnya. Keunggulan utama pendekatan ini adalah kemampuannya mengatasi masalah *cold-start* karena tidak memerlukan riwayat interaksi pengguna lain. CBF telah banyak digunakan secara efektif dalam berbagai domain berbasis teks, termasuk film, musik, buku, produk *e-commerce*, dan video game [2].

2.3. Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF adalah metode pembobotan statistik yang digunakan untuk merepresentasikan dokumen teks sebagai vektor numerik dalam ruang vektor multidimensi. Metode ini mengukur seberapa penting sebuah term terhadap suatu dokumen dalam konteks seluruh korpus data dengan mengombinasikan dua komponen utama. *Term Frequency* (TF) mengukur frekuensi relatif sebuah term dalam dokumen tertentu, sehingga term yang sering muncul dalam dokumen mendapat nilai TF tinggi. *Inverse Document Frequency* (IDF) mengukur kekhususan sebuah term di seluruh korpus, di mana term yang muncul di banyak dokumen mendapat nilai IDF rendah karena dianggap kurang diskriminatif, sedangkan term yang jarang muncul lintas dokumen mendapat nilai IDF tinggi. Nilai TF-IDF akhir diperoleh dengan mengalikan kedua komponen tersebut, sehingga term yang spesifik dan sering muncul dalam dokumen tertentu akan mendapatkan bobot tinggi [3].

2.4. Cosine Similarity

Cosine Similarity merupakan metode pengukuran kemiripan antara dua vektor dalam ruang multidimensi berdasarkan sudut kosinus yang terbentuk di antara keduanya. Nilai *Cosine Similarity* berkisar antara 0 hingga 1, di mana nilai 1 menunjukkan kedua vektor identik dan nilai 0 menunjukkan tidak ada kesamaan sama sekali. Keunggulan utama *Cosine Similarity* dibandingkan

metode jarak lain seperti *Euclidean Distance* adalah ketidakpekaan terhadap panjang dokumen, sehingga dokumen dengan panjang yang sangat berbeda tetap dapat dibandingkan secara adil. Dalam sistem rekomendasi, *Cosine Similarity* digunakan untuk membandingkan representasi vektor TF-IDF dari item guna menentukan item yang paling relevan bagi pengguna [7].

2.5. Penelitian Terdahulu

Sebuah sistem rekomendasi film berbasis CBF dengan TF-IDF menunjukkan bahwa pembobotan TF-IDF pada atribut sinopsis film mampu menghasilkan daftar rekomendasi dengan kemiripan konten yang tinggi [4]. Implementasi CBF pada sistem rekomendasi destinasi wisata juga menegaskan bahwa pendekatan berbasis konten sangat efektif pada kondisi *cold-start problem* [5]. Eksplorasi CBF dalam sistem rekomendasi event online menemukan bahwa kualitas tahap *preprocessing* teks secara signifikan memengaruhi akurasi sistem [6]. Kombinasi CBF dan *Cosine Similarity* pada sistem rekomendasi lagu membuktikan bahwa *Cosine Similarity* memberikan performa akurat dalam menghitung kemiripan antar vektor metadata [7]. Evaluasi khusus pada sistem rekomendasi game berbasis CBF dengan TF-IDF dan *Cosine Similarity* menjadi landasan langsung bagi penelitian ini [8].

Selanjutnya, implementasi sistem rekomendasi game pada platform Steam menggunakan dataset Kaggle menunjukkan bahwa kualitas dataset serta kelengkapan atribut metadata game berpengaruh signifikan terhadap relevansi hasil rekomendasi [9]. Penelitian lain mengenai kemiripan dokumen menggunakan representasi TF-IDF dan *Cosine Similarity* mengonfirmasi bahwa normalisasi vektor TF-IDF sebelum komputasi *Cosine Similarity* meningkatkan konsistensi hasil pengukuran secara signifikan [10]. Usulan pendekatan *Hybrid Recommendation System* untuk video game yang menggabungkan CBF dan *Collaborative Filtering* menyimpulkan bahwa pendekatan hibrida menghasilkan diversitas rekomendasi yang lebih baik daripada metode tunggal [11].

Terakhir, pengembangan sistem rekomendasi film berbasis CBF dengan pendekatan kemiripan deskripsi cerita menegaskan bahwa kualitas *text preprocessing* pada atribut deskripsi merupakan faktor kritis yang menentukan akurasi sistem rekomendasi berbasis teks [12].

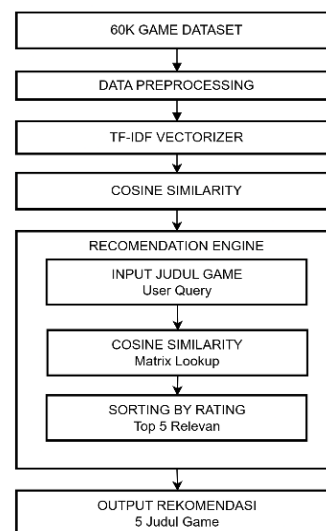
3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen laboratorium pada data sekunder. Implementasi sistem dilakukan sepenuhnya menggunakan bahasa pemrograman Python dengan memanfaatkan library *scikit-learn* pada lingkungan *Google Colab*. Spesifikasi lingkungan implementasi mencakup CPU Intel Xeon 2,2 GHz, RAM 12 GB, dan Python versi 3.10 dengan *scikit-*

learn versi 1.5. Dataset yang digunakan berupa 60.000 judul video game dari platform *Backloggd*, dengan atribut utama yang meliputi *title*, *summary*, *genres*, *platforms*, dan *rating*. Tahapan penelitian terdiri dari pengumpulan data, pra-proses data, representasi fitur, perhitungan kemiripan, dan evaluasi sistem.

3.1. Arsitektur Sistem

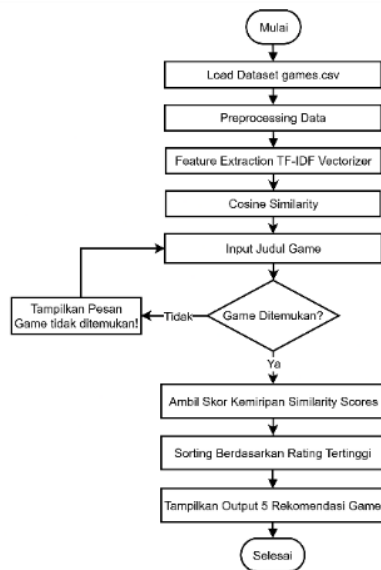
Arsitektur sistem rekomendasi yang dibangun terdiri dari empat lapisan utama yang bekerja secara berurutan. Lapisan pertama adalah tahap *data preprocessing* yang memproses data mentah menjadi data bersih dan terstandarisasi. Lapisan kedua adalah tahap ekstraksi fitur menggunakan TF-IDF Vectorizer yang mengubah data teks menjadi representasi vektor numerik. Lapisan ketiga adalah perhitungan *Cosine Similarity* yang menghasilkan matriks kemiripan antar game. Lapisan keempat adalah *recommendation engine* yang menerima input judul game dari pengguna, mencari skor kemiripan dalam matriks, lalu mengurutkan hasilnya berdasarkan *Rating* untuk menghasilkan 5 rekomendasi final. Arsitektur lengkap sistem ditunjukkan pada Gambar 1.



Gambar 1. Arsitektur Perancangan Sistem Rekomendasi Video Game

3.2. Alur Kerja Sistem (Flowchart)

Alur kerja sistem secara prosedural dimulai dari pemuatan dataset *games.csv*, dilanjutkan dengan tahap *preprocessing* dan ekstraksi fitur TF-IDF. Sistem kemudian menghitung matriks *Cosine Similarity* untuk seluruh pasangan game dalam dataset. Ketika pengguna memasukkan judul game sebagai input, sistem memeriksa keberadaan judul tersebut dalam dataset. Apabila judul tidak ditemukan, sistem menampilkan pesan "Game tidak ditemukan!". Apabila ditemukan, sistem mengambil skor kemiripan dari matriks, mengurutkan kandidat berdasarkan *Rating* tertinggi, dan menampilkan 5 rekomendasi game. Alur kerja lengkap sistem ditampilkan pada Gambar 2.



Gambar 2. Flowchart Sistem Rekomendasi Video Game

3.3. Pra-proses Data

Tahap *data cleaning* melibatkan serangkaian operasi pembersihan data yang mencakup: (1) penghapusan baris dengan nilai kosong (*missing value*) pada kolom *title* dan *summary*, (2) penghapusan data duplikat berdasarkan kolom *title*, (3) standarisasi format kolom *Genres* dan *Platforms* dari format *list string* menjadi teks datar, dan (4) penghapusan kolom yang tidak relevan seperti *Release_Date*, *Developers*, dan statistik interaksi pengguna. Proses ini mereduksi dataset dari 60.000 menjadi 37.669 data bersih, menghilangkan 37,2% data yang mengandung informasi tidak lengkap atau duplikat.

Tahap *text preprocessing* diterapkan pada kolom *Summary* dan mencakup tiga sub-tahap. *Case folding* mengonversi seluruh teks ke huruf kecil untuk memastikan konsistensi representasi. *Filtering* menghapus tanda baca, angka, karakter khusus, dan *stop words* bahasa Inggris menggunakan daftar *stop words* bawaan scikit-learn. *Tokenizing* memecah teks menjadi unit kata individual yang siap diproses. Setelah preprocessing, atribut *Summary*, *Genres*, dan *Platforms* dari setiap game digabungkan menjadi satu kolom *combined_features* yang menjadi representasi teks tunggal setiap game sebagai input TF-IDF.

3.4. Implementasi Algoritma TF-IDF

TF-IDF diimplementasikan menggunakan *TfidfVectorizer* dari library scikit-learn dengan formula sebagai berikut. *Term Frequency* dihitung sebagai:

$$TF(t, d) = \frac{f(t, d)}{\sum f(t', d)} \quad (1)$$

di mana $f(t, d)$ adalah jumlah kemunculan term t dalam dokumen d , dan penyebut adalah total seluruh term dalam d . *Inverse Document Frequency* dengan smoothing dihitung sebagai:

$$IDF(t) = \log \left(\frac{N + 1}{df(t) + 1} \right) + 1 \quad (2)$$

di mana N adalah total dokumen dalam korpus (37.669) dan $df(t)$ adalah jumlah dokumen yang mengandung term t . Nilai TF-IDF akhir diperoleh dari:

$$TF-IDF(t, d) = TF(t, d) \times IDF(t) \quad (3)$$

Implementasi menghasilkan matriks TF-IDF berukuran 37.669×71.951 , di mana setiap baris merepresentasikan satu game sebagai vektor berdimensi 71.951 (jumlah term unik dalam korpus). Term yang sangat spesifik terhadap satu game, seperti nama karakter atau nama dunia fiksi yang unik, mendapatkan bobot TF-IDF tertinggi dan berfungsi sebagai "sidik jari" konten game tersebut.

3.5. Perhitungan Cosine Similarity dan Sorting Rating

Cosine Similarity dihitung menggunakan fungsi *cosine_similarity* dari scikit-learn dengan formula:

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{|A| \times |B|} \quad (4)$$

Di mana A dan B adalah vektor TF-IDF dari dua game yang dibandingkan. Proses ini menghasilkan matriks simetris berukuran 37.669×37.669 yang menyimpan nilai kemiripan untuk setiap pasang game. Nilai kemiripan berkisar antara 0 (tidak ada kesamaan) hingga 1 (identik).

Mekanisme *sorting Rating* bekerja dalam dua tahap. Pertama, sistem mengambil top-20 kandidat game berdasarkan nilai *Cosine Similarity* tertinggi terhadap game input. Kedua, dari 20 kandidat tersebut, sistem mengurutkan ulang berdasarkan nilai *Rating* tertinggi dan mengambil 5 game teratas sebagai rekomendasi final. Pendekatan dua tahap ini memastikan bahwa rekomendasi yang dihasilkan tidak hanya relevan secara konten, tetapi juga berkualitas tinggi berdasarkan penilaian komunitas. Hal ini sejalan dengan temuan bahwa integrasi sinyal kualitas eksternal ke dalam sistem rekomendasi berbasis konten secara konsisten meningkatkan kepuasan pengguna terhadap hasil rekomendasi yang diberikan [11].

4. HASIL DAN PEMBAHASAN

4.1. Hasil Penelitian

Implementasi sistem menghasilkan beberapa output utama yang menjadi dasar evaluasi pada sub-bab berikutnya. Pertama, proses *preprocessing* berhasil mereduksi dataset dari 60.000 menjadi 37.669 game bersih, menghilangkan 37,2% data yang mengandung *missing value* atau duplikat. Kedua, *TF-IDF Vectorizer* menghasilkan matriks berukuran 37.669×71.951 , di mana 71.951 merupakan jumlah term unik yang diekstraksi dari seluruh kolom *combined_features*. Ketiga, komputasi *Cosine Similarity* menghasilkan matriks simetris 37.669×37.669 yang menyimpan nilai kemiripan konten untuk setiap pasangan game dalam dataset. Ringkasan output utama sistem ditampilkan pada Tabel 1.

Tabel 1. Ringkasan Output Utama Sistem

Komponen	Hasil
Jumlah data awal	60.000 judul game
Jumlah data bersih	37.669 judul game
Data dihapus	22.331 (37,2%)
Dimensi matriks TF-IDF	37.669×71.951
Dimensi matriks CS	37.669×37.669
Jumlah term unik	71.951 term
Output rekomendasi	Top-5 per query

Sistem berhasil dijalankan pada lingkungan *Google Colab* dengan waktu komputasi matriks TF-IDF kurang dari 10 detik dan komputasi matriks *Cosine Similarity* sekitar 45 detik untuk 37.669 game, menunjukkan performa komputasi yang efisien untuk dataset berskala besar. Setelah sistem dibangun, pengujian dilakukan menggunakan empat skenario lintas genre untuk mengevaluasi kualitas rekomendasi yang dihasilkan. Tabel 2 menampilkan contoh output rekomendasi sistem untuk setiap game input sebelum diterapkannya mekanisme *sorting Rating*.

Tabel 2. Contoh Output Rekomendasi Sistem (Sebelum Sorting Rating)

Game Input	Rekomendasi #1	Cosine Sim.
The Witcher 3	The Witcher 3: Blood and Wine	0,5812
FIFA 23	FIFA 22	0,4103
Minecraft	Minecraft: Story Mode	0,3947
GTA V	Grand Theft Auto IV	0,3521

4.2. Pengujian Sistem

Pengujian dilakukan pada empat skenario dengan game input dari genre yang berbeda-beda untuk mengevaluasi performa sistem secara lintas genre. Tabel 3 menampilkan profil keempat game yang digunakan sebagai input pengujian.

Tabel 3. Profil Game Input Pengujian Sistem

Judul Game	Genre	Rating
The Witcher 3: Wild Hunt	Adventure, RPG	4.2
FIFA 23	Simulator, Sport	3.3
Minecraft	Adventure, Simulator	4.1
Grand Theft Auto V	Adventure, Shooter	3.8

Dari Tabel 3 dapat dilihat bahwa keempat game input mencakup genre yang beragam, yaitu RPG, Sport, Simulator, dan Shooter, sehingga pengujian dapat mengevaluasi kemampuan generalisasi sistem secara komprehensif. Hasil pengujian menunjukkan bahwa sistem secara konsisten menghasilkan rekomendasi dari game-game dalam seri yang sama (intra-seri), yang mencerminkan kemampuan TF-IDF dalam mengenali kesamaan term naratif yang sangat spesifik pada satu waralaba game. Sebagai contoh, seluruh 5 rekomendasi untuk The Witcher 3 merupakan ekspansi dan edisi khusus game tersebut, sementara seluruh 5 rekomendasi untuk GTA V merupakan judul dalam seri Grand Theft Auto.

Satu anomali teridentifikasi pada skenario Minecraft, di mana game "Super Puyo Puyo 2" bergenre Puzzle masuk ke dalam daftar rekomendasi pada peringkat ke-4 meskipun tidak memiliki kesamaan genre dengan Minecraft. Anomali ini terjadi karena deskripsi game tersebut mengandung term "blocks" yang memiliki bobot TF-IDF tinggi dan juga muncul menonjol dalam deskripsi Minecraft. Kasus ini secara langsung mencerminkan keterbatasan fundamental pendekatan TF-IDF yang tidak mampu membedakan konteks semantik penggunaan sebuah term, melainkan hanya mengandalkan kesamaan term secara literal.

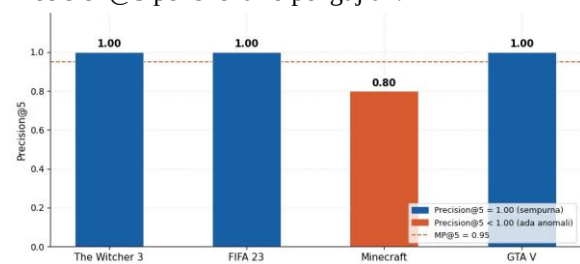
4.3. Evaluasi dengan Presicion@5

Evaluasi sistem menggunakan metrik *Precision at K* (*Precision@K*) dengan nilai $K = 5$, yang merupakan metrik standar dalam evaluasi sistem rekomendasi. *Precision@K* mengukur proporsi item relevan di antara K item teratas yang direkomendasikan oleh sistem, dengan formula sebagai berikut:

$$\text{Precision@K} = \frac{|\{\text{item relevan}\} \cap \{\text{top-K}\}|}{K} \quad (5)$$

$$\text{Mean Precision@K (MP@K)} = \frac{1}{N} \sum_{i=1}^N \text{Precision@K}(i) \quad (6)$$

Kriteria relevansi yang digunakan adalah kesamaan genre: sebuah game rekomendasi dinyatakan relevan apabila memiliki setidaknya satu genre yang sama dengan game input. Pendekatan ini memberikan toleransi terhadap perbedaan genre sekunder selama genre utama tetap terwakili dalam rekomendasi. Gambar 3 menampilkan visualisasi nilai *Precision@5* per skenario pengujian.



Gambar 3. Nilai Precision@5 per Skenario Pengujian

Dari Gambar 3 dapat dilihat bahwa tiga dari empat skenario pengujian (The Witcher 3, FIFA 23, dan GTA V) mencapai nilai *Precision@5* = 1,00 yang berarti seluruh 5 rekomendasi yang dihasilkan relevan secara genre. Skenario Minecraft mencapai *Precision@5* = 0,80 akibat satu anomali yang telah dijelaskan sebelumnya. *Mean Precision@5* dari keseluruhan empat skenario adalah 0,95, yang merupakan nilai yang sangat tinggi dan mengonfirmasi bahwa sistem CBF berbasis TF-IDF dan *Cosine Similarity* ini secara umum sangat efektif dalam menghasilkan rekomendasi yang relevan secara konten pada domain video game.

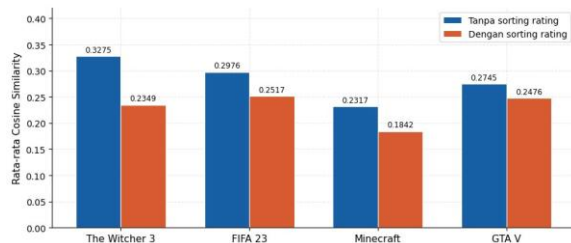
Tabel 4. Hasil Perhitungan Precision@5 per Skenario

Game Input	Relevan (K=5)	Tidak Relevan	Precision@5
The Witcher 3	5	0	1,00
FIFA 23	5	0	1,00
Minecraft	4	1	0,80
GTA V	5	0	1,00
Mean Precision@5			0,95

Dari Tabel 4 dapat disimpulkan bahwa nilai MP@5 = 0,95 merupakan hasil yang sangat memuaskan untuk sistem rekomendasi berbasis teks pada skala dataset yang besar (37.669 game). Nilai ini mengonfirmasi bahwa penggabungan tiga fitur tekstual (*Summary*, *Genres*, *Platforms*) dalam satu representasi *combined_features* berhasil menciptakan profil konten yang kaya dan diskriminatif untuk setiap game.

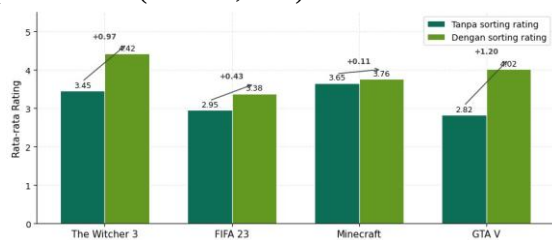
4.4. Analisis Pengaruh Sorting Rating

Untuk menganalisis pengaruh mekanisme *sorting Rating* secara kuantitatif, dilakukan perbandingan komposisi rekomendasi sebelum dan sesudah penerapan *sorting* pada keempat skenario pengujian. Gambar 4 menampilkan perbandingan rata-rata *Cosine Similarity* antara kondisi tanpa *sorting* dan dengan *sorting Rating*.



Gambar 4. Perbandingan Rata-rata Cosine Similarity: Tanpa vs Dengan Sorting Rating

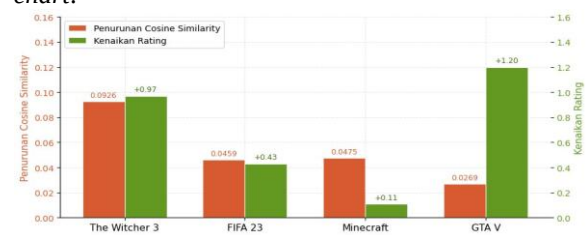
Dari Gambar 4 terlihat bahwa pada seluruh skenario, bar biru (tanpa *sorting*) selalu lebih tinggi dibandingkan bar oranye (dengan *sorting*). Hal ini berarti penerapan *sorting Rating* memang menyebabkan sedikit penurunan rata-rata *Cosine Similarity* pada hasil akhir, karena beberapa game dengan similarity tertinggi digantikan oleh game dengan rating yang lebih tinggi. Penurunan terbesar terjadi pada The Witcher 3 (selisih 0,0926) dan terkecil pada GTA V (selisih 0,0269).



Gambar 5. Perbandingan Rata-rata Rating: Tanpa vs Dengan Sorting Rating

Gambar 5 menampilkan perbandingan rata-rata *Rating* sebelum dan sesudah penerapan *sorting*, yang merupakan sisi lain dari trade-off tersebut, sebaliknya terlihat bahwa bar hijau muda (dengan *sorting*) selalu lebih tinggi dibandingkan bar hijau tua (tanpa *sorting*). Kenaikan *Rating* terbesar terjadi pada GTA V (+1,20 poin, dari 2,82 menjadi 4,02), diikuti The Witcher 3 (+0,97), FIFA 23 (+0,43), dan Minecraft (+0,11). Kenaikan terkecil pada Minecraft terjadi karena kandidat top-20 berdasarkan similarity sudah didominasi oleh versi-versi Minecraft yang memiliki rating tinggi, sehingga *sorting Rating* tidak banyak mengubah komposisi.

Untuk memvisualisasikan kedua aspek trade-off secara bersamaan, Gambar 6 menyajikan perbandingan besaran penurunan *Cosine Similarity* dan kenaikan *Rating* menggunakan *dual-axis bar chart*.



Gambar 6. Trade-off: Penurunan Cosine Similarity vs Kenaikan Rating Setelah Sorting

Dari Gambar 6, GTA V menunjukkan profil trade-off paling menguntungkan: penurunan *Cosine Similarity* sangat kecil (0,0269, bar oranye pendek) namun kenaikan *Rating* sangat besar (+1,20, bar hijau tertinggi). Sebaliknya, The Witcher 3 menunjukkan penurunan *Cosine Similarity* terbesar (0,0926) meskipun kenaikan *Rating*-nya juga signifikan (+0,97). Temuan ini penting karena menunjukkan bahwa besarnya manfaat *sorting Rating* bergantung pada distribusi *Rating* dalam populasi kandidat game yang serupa.

Tabel 5. Perbandingan Rata-rata Cosine Similarity dan Rating

Game Input	CS Tanpa Sort	CS Dg Sort	Rating Tanpa Sort	Rating Dg Sort
The Witcher 3	0,3275	0,2349	3,45	4,42
FIFA 23	0,2976	0,2517	2,95	3,38
Minecraft	0,2317	0,1842	3,65	3,76
GTA V	0,2745	0,2476	2,82	4,02
Rata-rata	0,2828	0,2296	3,22	3,90

Dari Tabel 5 terlihat secara keseluruhan bahwa *sorting Rating* menyebabkan penurunan rata-rata *Cosine Similarity* sebesar 0,0532 (dari 0,2828 menjadi 0,2296) dan kenaikan rata-rata *Rating* sebesar 0,68 poin (dari 3,22 menjadi 3,90). Peningkatan kualitas yang diperoleh jauh lebih signifikan dibandingkan pengurangan kemiripan yang terjadi, sehingga mekanisme *sorting Rating* yang diterapkan memberikan nilai tambah nyata pada kualitas

rekomendasi akhir. Temuan ini mengonfirmasi bahwa penggunaan *Cosine Similarity* sebagai filter utama dan *Rating* sebagai filter sekunder menghasilkan keseimbangan yang optimal antara relevansi konten dan kualitas penilaian komunitas.

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengimplementasikan sistem rekomendasi video game berbasis *Content-Based Filtering* menggunakan algoritma TF-IDF dan *Cosine Similarity* pada dataset 37.669 game dari platform *Backloggd*. Sistem menghasilkan matriks TF-IDF berukuran 37.669×71.951 yang merepresentasikan seluruh game sebagai vektor teks terbobot. Evaluasi menggunakan $Precision@5$ menghasilkan $Mean\ Precision@5 = 0,95$, yang menunjukkan efektivitas tinggi dalam menghasilkan rekomendasi yang relevan secara konten lintas genre. Mekanisme *sorting Rating* terbukti secara konsisten meningkatkan rata-rata *Rating* rekomendasi dari 3,22 menjadi 3,90 dengan trade-off penurunan *Cosine Similarity* yang terukur namun dapat diterima (rata-rata 0,0532). Sistem ini efektif mengatasi masalah *cold-start* karena tidak memerlukan data historis pengguna. Keterbatasan utama sistem adalah ketidakmampuan TF-IDF menangkap konteks semantik, yang dapat menyebabkan anomali rekomendasi seperti pada skenario *Minecraft*. Untuk pengembangan selanjutnya, disarankan penggunaan embedding semantik seperti Word2Vec atau BERT untuk meningkatkan pemahaman konteks, penambahan fitur gameplay abstrak, pembaruan dataset secara berkala melalui API platform, penanganan *missing rating* melalui imputasi berbasis konten, serta pengembangan *Hybrid Filtering* untuk meningkatkan personalisasi rekomendasi.

DAFTAR PUSTAKA

- [1] S. I. Adam and W. G. Mokodaser, "Implementasi sistem rekomendasi produk e-commerce menggunakan content-based filtering berbasis cosine similarity," *J. Sist. Inf. dan Tek. Komput.*, 2024, [Online]. Available: <https://ejournal.caturisakti.ac.id/index.php/simtek/article/view/1665>
- [2] T. Ridwansyah, B. Subartini, and S. Sylviani, "Penerapan Metode Content-Based Filtering pada Sistem Rekomendasi," *Math. Sci. Appl. J.*, vol. 4, no. 2, pp. 70–77, Apr. 2024, doi: 10.22437/MSA.V4I2.32136.
- [3] N. O. Idris and F. Pontoioyo, "Sistem rekomendasi produk makeup berbasis content-based filtering dengan TF-IDF dan cosine similarity," *KETIK J. Inform.*, 2025, [Online]. Available: <https://jurnal.faatuatua.com/index.php/KETIK/article/view/311>
- [4] A. Febrian and E. D. Permana, "Sistem rekomendasi film menggunakan metode content-based filtering dengan algoritma TF-IDF," *Al-Aqlu J. Ilmu Agama, Sos. dan Budaya*, vol. 3, no. 1, 2024, [Online]. Available: <https://jurnal.yalamqa.com/index.php/aqlu/article/view/494>
- [5] S. O. Malikussaleh, D. Abdullah, and A. Razi, "Sistem rekomendasi destinasi wisata menggunakan content-based filtering," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 10, no. 1, 2025, [Online]. Available: <https://jurnal.univrab.ac.id/index.php/rabit/article/view/6532>
- [6] F. Nurfalah, Asriyanik, and A. Pambudi, "Sistem rekomendasi event online menggunakan metode content-based filtering," *J. ELKOM (Elektronika dan Komputer)*, vol. 15, no. 2, 2022, [Online]. Available: <https://journal.stekom.ac.id/index.php/elkom/article/view/736>
- [7] A. Y. Timur, "Sistem rekomendasi lagu menggunakan metode content-based filtering dan cosine similarity," *JITET (Jurnal Inform. Terap. dan Edukasi Terintegrasi)*, vol. 12, no. 1, 2024, [Online]. Available: <https://journal.eng.unila.ac.id/index.php/jitet/article/view/5949>
- [8] M. Farhan, M. R. Andreawan, R. Antonius, and N. D. Ulhag, "Evaluasi dan Pengembangan Sistem Rekomendasi Game Berbasis Content-Based Filtering dengan TF-IDF dan Cosine Similarity," *Biner J. Ilmu Komputer, Tek. dan Multimed.*, vol. 2, no. 4, pp. 400–408, 2024, [Online]. Available: <https://journal.mediapublikasi.id/index.php/Biner/article/view/4638>
- [9] T. A. Nabila, L. H. Angwarmasse, D. K. Cahyaningrum, and B. V. Christioko, "Sistem Rekomendasi Game pada Steam dengan Analisis Dataset Kaggle," in *Seminar Nasional Inovasi dan Tren Teknologi Informasi (SINATTI) 2025*, Semarang: Universitas Semarang, 2025. [Online]. Available: <https://conference.usm.ac.id/index.php/sinatti/sinatti2025/paper/view/1203>
- [10] A. Widiyanto, E. Pebriyanto, F. Fitriyanti, and M. Marna, "Document Similarity Using Term Frequency-Inverse Document Frequency Representation and Cosine Similarity," *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 4, no. 2, pp. 149–153, Aug. 2024, doi: 10.20895/DINDA.V4I2.1589.
- [11] A. Manikantan, "A Hybrid Recommendation System for Video Games: Combining Content-based & Collaborative Filtering," *IJRASET*, vol. 9, 2021, doi: 10.22214/ijraset.2021.38246.
- [12] R. M. Harahap and A. N. Rachman, "SISTEM REKOMEDASI FILM BERDASARKAN KEMIRIPAN DESKRIPSI CERITA MENGGUNAKAN METODE CONTENT BASED FILTERING," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3, pp. 2830–2837, Jul. 2025, doi: 10.23960/JITET.V13I3.6577.