AITI: Jurnal Teknologi Informasi, Volume 23 No. 2 Juni 2026, 319-333
ISSN 1693-8348 E-ISSN 2615-7128
DOI: 10.24246/aiti.v23.i2.319-333

JURNAL TEKNOLOGI INFORMASI

Analisis dan implementasi *Content-Based Filtering* dengan *Cosine Similarity* untuk sistem rekomendasi tugas akhir mahasiswa

Ero Wahyu Pratomo*, Ema Utami
Universitas Amikom Yogyakarta
Jl. Ring Road Utara, Ngringin, Condongcatur, Kec. Depok,
Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281
Email: *erowahyu@students.amikom.ac.id

Riwayat Artikel

Diterima:	Direvisi:	Disetujui:
12-06-2025	28-08-2025	29-08-2024

Abstract

This study aims to analyze and implement a student thesis recommendation system based on the Content-Based Filtering approach, using the Cosine Similarity method. The proposed system is designed to assist students in identifying relevant thesis topics based on their interests and prior research. The methods used in this research include data collection from a repository through the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), text preprocessing by removing Indonesian stop words, the formation of TF-IDF matrices, and the calculation of document similarity using cosine similarity. The experimental results show that the system successfully provides relevant recommendations with satisfactory precision and recall, as well as fast response times, making it suitable for real-world applications. These findings suggest that the Content-Based Filtering approach, combined with cosine similarity, is effective at capturing content similarity in student theses and providing recommendations that align with user needs. Therefore, the developed system can serve as an efficient and accurate solution to support students in selecting appropriate thesis topics and also facilitate access to relevant academic references.

Keywords: *recommendation system, Content-Based Filtering, Cosine Similarity, text preprocessing*

Abstrak

Penelitian ini bertujuan untuk menganalisis dan mengimplementasikan sistem rekomendasi tugas akhir mahasiswa berbasis *Content-Based Filtering* dengan memanfaatkan metode *Cosine Similarity*. Sistem yang dikembangkan dirancang untuk membantu mahasiswa menemukan topik tugas akhir yang relevan berdasarkan minat dan referensi penelitian sebelumnya. Metode yang digunakan dalam penelitian ini meliputi pengumpulan data dari repository melalui *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH), *preprocessing* teks dengan menghapus *stop words* dalam bahasa Indonesia, pembentukan matriks TF-IDF, serta perhitungan kemiripan antardokumen menggunakan *cosine similarity*. Hasil eksperimen menunjukkan bahwa sistem ini berhasil memberikan rekomendasi yang relevan dengan nilai precision dan recall yang memadai, serta waktu respons yang cepat, sehingga layak diterapkan dalam aplikasi nyata. Temuan ini menunjukkan bahwa pendekatan *Content-Based Filtering* dengan *cosine similarity* efektif dalam menangkap kemiripan konten tugas akhir mahasiswa serta memberikan rekomendasi yang sesuai dengan kebutuhan pengguna. Dengan demikian, sistem yang

dikembangkan dapat menjadi solusi yang efisien dan akurat untuk membantu mahasiswa menentukan topik tugas akhir yang tepat, sekaligus mempermudah akses terhadap referensi akademik yang relevan.

Kata kunci: sistem rekomendasi, *Content-Based Filtering, Cosine Similarity*, *preprocessing* teks

Pendahuluan

Sistem rekomendasi kini memegang peranan krusial di berbagai sektor, termasuk dalam ranah akademik. Dalam konteks pendidikan tinggi, pemilihan topik tugas akhir merupakan langkah krusial yang dapat memengaruhi keberhasilan akademik dan pengembangan karir mahasiswa. Tantangan seperti banyaknya pilihan topik, minimnya informasi yang terstruktur, dan kurangnya pemahaman terhadap preferensi individu menjadikan pengembangan sistem rekomendasi yang efektif sangat penting untuk mendukung pengambilan keputusan mahasiswa. Selain itu, sistem repositori institusi sering kali hanya menyediakan pencarian berbasis kata kunci sederhana, yang kurang mampu menyajikan hasil yang relevan secara semantik [1], [2]. Masalah ini menjadi lebih kompleks seiring pertumbuhan pesat koleksi digital, yang menyebabkan pencarian manual menjadi tidak efisien dan rawan terhadap informasi yang terlewat [1].

Metode yang biasa digunakan dalam sistem rekomendasi adalah *content-based filtering*, yang menganalisis atribut atau fitur dari item, deskripsi atau kata kunci topik tugas akhir. Sistem ini memberikan rekomendasi berdasarkan kemiripan antara fitur item yang diminati pengguna dan item lain dalam dataset. Kelebihannya terletak pada kemampuan untuk menghasilkan rekomendasi personal tanpa bergantung pada data pengguna lain, sehingga cocok diterapkan pada sistem akademik yang sering kali memiliki keragaman konten [3], [4].

Untuk mengukur kemiripan antar item, digunakan metode *cosine similarity*, yang menghitung sudut antar vektor dalam ruang multidimensi berdasarkan bobot kata dari teknik TF-IDF. Nilai kesamaan dapat berada dalam rentang -1 hingga 1, dengan nilai yang mendekati 1 menunjukkan tingkat kesamaan yang tinggi. Pendekatan ini terbukti efisien untuk menangani data teks dan numerik, serta mampu menggambarkan hubungan semantik antar dokumen secara akurat [1], [3], [5], [6], [7]. Sejumlah studi juga menyoroti bahwa representasi TF-IDF sangat berperan dalam meningkatkan kualitas rekomendasi berbasis konten [2], [3], [8].

Di samping *content-based filtering*, terdapat pula *collaborative filtering*, yang mengandalkan data interaksi pengguna seperti rating. Pendekatan ini efektif dalam mengungkap preferensi kolektif, namun rentan terhadap masalah *cold start* dan *sparsity* ketika data pengguna terbatas. Dalam hal ini, *cosine similarity* juga diterapkan untuk mengukur tingkat kesamaan antara pengguna atau item berdasarkan pola interaksi yang terjadi [6], [7], [9], [10], [11]. Meskipun memiliki potensi yang tinggi, pendekatan ini menuntut data pengguna yang lengkap dan dinamis, yang tidak selalu tersedia dalam sistem akademik [1].

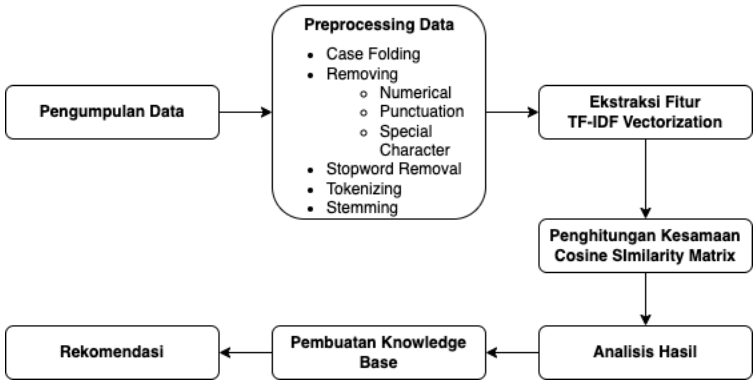
Untuk mengatasi keterbatasan tersebut, dikembangkan *hybrid filtering*, yang menggabungkan *content-based* dan *collaborative filtering*. Pendekatan ini memanfaatkan keunggulan kedua metode, di mana *cosine similarity* digunakan dalam analisis konten, sementara *collaborative filtering* diterapkan melalui teknik seperti *matrix factorization* [2], [3], [12]. Selain itu, penelitian juga menunjukkan bahwa *cosine similarity* dapat diterapkan dalam konteks lain seperti analisis afiliasi *marketplace* dan segmentasi visual berbasis atribut, membuktikan fleksibilitas teknik ini dalam pengolahan data yang kompleks [1], [13]. Dengan demikian, pemilihan metode rekomendasi perlu mempertimbangkan ketersediaan data dan tujuan aplikasi, di mana *content-based filtering* tetap menjadi opsi kuat untuk sistem rekomendasi tugas akhir berbasis metadata.

Metode Penelitian

Implementasi *content-based filtering* dilakukan setelah *metadata* tugas akhir diperoleh, untuk merekomendasikan topik tugas akhir berdasarkan kemiripan fitur. *Content-based filtering* dalam penelitian ini menggunakan representasi fitur berbasis teks yang diekstraksi dari metadata, seperti judul, abstrak, dan kata kunci. Fitur-fitur ini kemudian diolah menjadi vektor numerik menggunakan teknik seperti *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk mengekspresikan pentingnya setiap kata dalam konteks dokumen dan koleksi data [3], [6]. *Cosine similarity* diterapkan sebagai metode utama untuk menentukan tingkat kesamaan antara dua vektor fitur. Secara matematis, metode ini mengukur sudut kosinus antara dua vektor dalam ruang berdimensi tinggi, dengan nilai yang berkisar antara -1 hingga 1. Nilai 1 menunjukkan vektor yang identik atau sangat mirip, sedangkan nilai 0 menunjukkan vektor yang orthogonal atau tidak memiliki kemiripan. Rumus *cosine similarity* dapat dilihat pada Persamaan (1) [5].

$$Cosine\ Similarity = \frac{A \cdot B}{\|A\|\ x\ \|B\|} \tag{1}$$

di mana *A* dan *B* adalah vektor fitur dari dua item (topik tugas akhir), dan *n* adalah jumlah dimensi fitur. Proses ini memungkinkan sistem untuk menghitung kesamaan antar topik berdasarkan representasi fitur mereka secara kuantitatif [3]. Dalam sistem rekomendasi yang dikembangkan, setiap topik tugas akhir direpresentasikan sebagai vektor TF-IDF yang mencerminkan bobot kata-kata penting dalam metadata. Ketika seorang mahasiswa memasukkan preferensi atau memilih topik awal, sistem akan menghitung *cosine similarity* antara vektor topik tersebut dengan seluruh vektor topik lain dalam dataset. Topik-topik dengan nilai *cosine similarity* tertinggi akan direkomendasikan sebagai alternatif yang relevan dan sesuai dengan minat mahasiswa [14], [15].



Gambar 1 Diagram penelitian

Berikut ini adalah penjelasan lengkap dari diagram alur sistem rekomendasi tugas akhir mahasiswa berbasis *Content-Based Filtering* menggunakan *Cosine Similarity* yang ditampilkan dalam Gambar 1.

1. **Pengumpulan Data**
   Tahapan awal dimulai dari proses pengumpulan data. Data ini diambil dari repositori akademik, khususnya metadata tugas akhir mahasiswa yang tersedia melalui protokol *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH). Metadata yang dikumpulkan biasanya mencakup informasi penting seperti judul tugas akhir (*title*), abstrak atau deskripsi (*description*), nama penulis (*creator*), tanggal publikasi (*date*), dan tautan ke dokumen (*relation*). Data tersebut tidak digunakan secara langsung, melainkan dijadikan bahan mentah yang akan diproses lebih lanjut untuk menghasilkan rekomendasi yang berkualitas dan relevan.
2. ***Preprocessing* Data**
   Setelah data berhasil dikumpulkan, langkah selanjutnya yaitu *preprocessing data*. Tahap ini krusial karena data teks yang diperoleh dari metadata biasanya belum bersih dan seragam. *Preprocessing* dilakukan untuk meningkatkan kualitas data dan menyederhanakan teks agar lebih mudah diolah oleh mesin. Langkah-langkah *preprocessing* dalam diagram mencakup:
   a. ***Case Folding***: Mengubah semua huruf dalam teks menjadi huruf kecil. Hal ini berguna untuk menghindari redudansi akibat perbedaan kapitalisasi, seperti kata "Mahasiswa" dan "mahasiswa" yang secara makna sama.
   b. ***Removing***: Proses ini menghapus elemen-elemen yang dianggap tidak penting dalam analisis teks, yaitu: *Numerical*, *Punctuation* dan *Special Character*
   c. ***Stopword Removal***: Kosakata yang sering muncul namun tidak memberi makna penting dalam konteks, seperti “dan”, “yang”, “dari”, akan dihapus.
   d. ***Tokenizing***: Teks dipecah menjadi unit kata individual yang disebut token. Misalnya, kalimat “Sistem rekomendasi tugas akhir” akan diubah menjadi tiga token: “sistem”, “rekomendasi”, dan “tugas akhir”.
   e. ***Stemming***: Setiap token dikembalikan ke bentuk dasarnya. Sebagai contoh, “pengajaran”, “mengajar”, dan “pengajar” semuanya di-*stem* menjadi

bentuk dasar “ajar”. Ini bertujuan agar sistem mengenali kata yang berbeda tapi bermakna sama sebagai satu entitas.

Langkah-langkah ini secara keseluruhan memastikan bahwa teks yang akan dianalisis benar-benar bersih, ringkas, dan representatif.

3. **Ekstraksi Fitur – TF-IDF *Vectorization***

   Setelah teks dibersihkan, dokumen-dokumen tersebut kemudian diubah menjadi representasi numerik menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). TF-IDF menghitung bobot setiap kata dalam dokumen berdasarkan frekuensinya relatif terhadap keseluruhan koleksi dokumen. Kata yang sering muncul dalam satu dokumen namun jarang terdapat di dokumen lain akan memperoleh bobot yang lebih tinggi, menunjukkan bahwa kata tersebut memiliki kekhususan dalam membedakan isi dokumen tersebut. Proses ini menghasilkan matriks TF-IDF yang digunakan sebagai dasar untuk mengukur kemiripan antar dokumen.

4. **Penghitungan Kesamaan – *Cosine Similarity Matrix***

   Setelah representasi dokumen dalam bentuk vektor numerik diperoleh, sistem mulai menghitung tingkat kemiripan antar dokumen dengan menggunakan metode *Cosine Similarity*. *Cosine similarity* mengukur sudut kosinus antara dua vektor di ruang multidimensi. Jika dua vektor menuju ke arah yang sama, nilai kemiripan mendekati 1 (artinya kontennya sangat mirip). Sebaliknya, jika kedua vektor tegak lurus atau menuju ke arah yang berbeda, nilai kemiripan mendekati 0 (artinya tidak mirip). Proses ini menghasilkan matriks kemiripan, di mana setiap nilai di dalamnya menunjukkan seberapa mirip satu dokumen dengan dokumen lainnya.

5. **Analisis Hasil**

   Setelah diperoleh nilai kemiripan antar dokumen, dilakukan analisis untuk mengevaluasi hasil tersebut. Analisis ini bertujuan memastikan bahwa dokumen yang memiliki skor kemiripan tinggi memang memiliki keterkaitan secara substansi. Selain itu, analisis ini digunakan untuk menilai efektivitas metode *cosine similarity* dalam mengenali hubungan antar topik, serta untuk mengidentifikasi adanya pola atau kelompok dokumen yang cenderung memiliki kesamaan konten. Tahapan ini menjadi bagian penting dari proses evaluasi awal sebelum sistem diterapkan secara aktif dalam memberikan rekomendasi kepada pengguna.

6. **Pembuatan *Knowledge Base***

   Informasi yang dihasilkan dari proses representasi TF-IDF dan perhitungan *cosine similarity* disimpan dalam sebuah *knowledge base*. Basis pengetahuan ini berisi matriks TF-IDF yang merepresentasikan dokumen dalam bentuk numerik, serta matriks *cosine similarity* yang mencatat tingkat kemiripan antar dokumen. Dengan menyimpan kedua matriks ini, sistem dapat menghindari perhitungan ulang setiap kali pengguna mengajukan permintaan rekomendasi, sehingga proses pencarian menjadi lebih cepat, efisien, dan responif terhadap kebutuhan pengguna.

7. Rekomendasi

Tahap akhir dari sistem adalah proses pemberian rekomendasi. Pada tahap ini, ketika pengguna memasukkan *input* berupa judul atau abstrak tugas akhir, sistem akan mencocokkannya dengan dokumen-dokumen yang telah dianalisis sebelumnya. Kemudian, sistem menghitung nilai kemiripan menggunakan cosine similarity dan mengambil dokumen-dokumen dengan skor tertinggi. Untuk menjaga relevansi hasil, sistem menerapkan ambang batas (*threshold*) tertentu guna menyaring dokumen yang kurang mirip. Akhirnya, sistem menampilkan sejumlah hasil rekomendasi teratas (*Top-N*) yang paling mendekati *input* pengguna, sehingga rekomendasi yang diberikan lebih tepat dan sesuai dengan kebutuhan. Hasil rekomendasi ini diharapkan memberikan informasi yang relevan dan membantu pengguna menemukan tugas akhir yang memiliki kesamaan konten atau topik secara substansial.

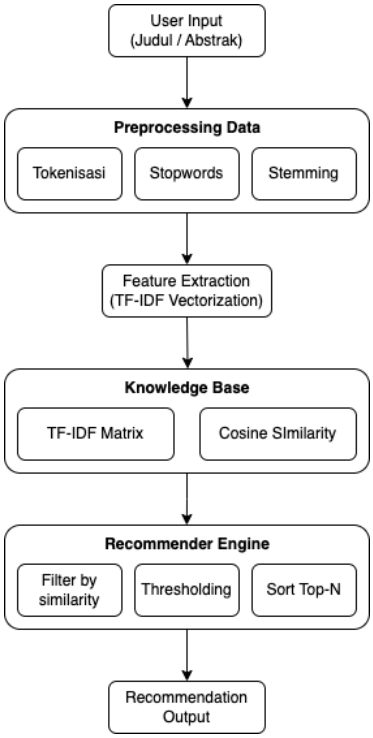
Diagram menggambarkan alur sistem rekomendasi tugas akhir secara runtut dan terintegrasi, mulai dari pengumpulan data hingga pemberian rekomendasi. Setiap tahap memiliki peran yang signifikan dalam menghasilkan rekomendasi yang relevan dan berguna bagi pengguna. Dengan menggunakan pendekatan *content-based filtering* dan *cosine similarity*, sistem dapat memberikan rekomendasi yang dipersonalisasi berdasarkan kesamaan isi dokumen. Representasi numerik dokumen dilakukan melalui TF-IDF, sementara *cosine similarity* digunakan untuk mengukur tingkat kemiripan antar dokumen.

Pendekatan ini didukung oleh berbagai studi yang membuktikan efektivitas *cosine similarity* dalam mengukur kemiripan teks, khususnya pada metadata akademik. OAI-PMH juga memperkuat sistem melalui penyediaan data yang terstruktur dan representatif. Sistem ini tidak hanya relevan untuk rekomendasi topik, tetapi juga fleksibel untuk dikembangkan lebih lanjut, seperti dengan penambahan fitur *filtering*, penggunaan *word embedding*, atau integrasi ke sistem akademik yang lebih luas.

Hasil dan Pembahasan

A. Arsitektur Sistem

Dari hasil analisis tersebut akan dibuat sebuah *knowledge base* untuk sistem rekomendasi tugas akhir seperti Gambar 2. Gambar 2 menggambarkan arsitektur sistem rekomendasi tugas akhir mahasiswa yang dirancang modular untuk menjamin efisiensi, akurasi, dan skalabilitas. Proses dimulai dengan akuisisi metadata seperti judul, deskripsi, pembuat, tanggal, dan tautan dari repositori institusi melalui protokol OAI-PMH. Selanjutnya, data diproses melalui tahapan *preprocessing* yang mencakup konversi huruf ke huruf kecil, penghapusan tanda baca, angka, dan *stopword*, serta normalisasi kata untuk membentuk format teks yang bersih dan konsisten. Selanjutnya, dilakukan ekstraksi fitur dengan menggunakan metode TF-IDF untuk mengonversi teks menjadi vektor numerik yang kemudian disimpan dalam basis pengetahuan bersama *metadata* aslinya.



Gambar 2 Arsitektur sistem

Kemiripan antar dokumen dihitung menggunakan *cosine similarity*, yang mengukur sudut antar vektor dalam ruang multidimensi, menghasilkan nilai antara -1 hingga 1 untuk merepresentasikan tingkat kesamaan semantik [3], [5]. Modul rekomendasi akan mencocokkan *input* pengguna terhadap seluruh data menggunakan *threshold* kemiripan (≥ 0,3), dan menyusun hasil rekomendasi berdasarkan skor tertinggi. Arsitektur ini juga dilengkapi dengan modul evaluasi dan monitoring performa menggunakan metrik seperti *precision, recall*, dan waktu respon rata-rata. Hasil evaluasi divisualisasikan melalui *confusion matrix* dan histogram distribusi kemiripan untuk menilai akurasi dan efisiensi sistem secara menyeluruh [13], [16].

B. Skenario Eksperimen

Untuk mengevaluasi performa sistem rekomendasi tugas akhir berbasis *content-based filtering* dengan pendekatan *cosine similarity*, dilakukan dua skenario eksperimen dengan jumlah dan cakupan data yang berbeda yaitu: (1) Skenario pertama menggunakan 1849 data tugas akhir dari tahun 2024 dan (2) Skenario kedua menggunakan seluruh data tanpa batasan tahun, sebanyak 28.803 entri dari repositori eprints.amikom.ac.id.

Kedua eksperimen menggunakan *input* judul yang sama, yaitu "TEKNIK PERBANDINGAN MACHINE LEARNING UNTUK PENYAKIT DIAGNOSIS DAN PREDIKSI JANTUNG", untuk menguji kualitas rekomendasi. Gambar 3 dan Gambar 4 menunjukkan hasil skenario 1 dan 2.

```
Average Precision@5: 0.6004
Average Recall@5: 0.5394
Measuring response time for recommendations on test data...
Average response time per recommendation: 0.0021 seconds
Masukkan judul atau deskripsi tugas akhir untuk rekomendasi: TEKNIK PERBANDINGAN MACHINE LEARNING UNTUK PENYAKIT DIAGNOSIS DAN PREDIKSI JANTUNG
Rekomendasi berdasarkan input Anda (threshold=0.3):
- Title: PERBANDINGAN TEKNIK MACHINE LEARNING UNTUK DIAGNOSIS DAN PREDIKSI PENYAKIT JANTUNG
  Link: https://eprints.amikom.ac.id//id/eprint/26265/
  Creator: Akbar, Muhammad Ilham
  Date: 2024-03-19
  Similarity Score: 1.0000

- Title: {JALUR SCIENTIST} DIAGNOSIS DAN PREDIKSI DIABETES MENGGUNAKAN TEKNIK DATA MINING DAN MACHINE LEARNING
  Link: https://eprints.amikom.ac.id//id/eprint/25285/
  Creator: Fadillah, Muhammad Iqbal
  Date: 2024-01-23
  Similarity Score: 0.5049

- Title: {JALUR SCIENTIST} PENERAPAN ALGORITMA MACHINE LEARNING UNTUK
SISTEM PREDIKSI PENYAKIT OSTEOPOROSIS
  Link: https://eprints.amikom.ac.id//id/eprint/28687/
  Creator: Sujana, Rajendra Artanto Wiryawan
  Date: 2024-11-20
  Similarity Score: 0.4538

- Title: OPTIMASI HYPERPARAMETER GRIDSEARCHCV PADA MACHINE LEARNING UNTUK KLASIFIKASI PENYAKIT JANTUNG
  Link: https://eprints.amikom.ac.id//id/eprint/28779/
  Creator: Jannah, Septa Miftakhul
  Date: 2024-12-23
  Similarity Score: 0.3997

- Title: OPTIMASI ALGORITMA MACHINE LEARNING DENGAN HYPERPARAMETER TUNING GRID SEARCH CV PADA KLASIFIKASI PENYAKIT JANTUNG
  Link: https://eprints.amikom.ac.id//id/eprint/28825/
  Creator: Hakim, Rahman
  Date: 2024-12-18
  Similarity Score: 0.3921
```

Gambar 3 Hasil skenario 1

```
Average Precision@5: 0.8827
Average Recall@5: 0.5022
Measuring response time for recommendations on test data...
Average response time per recommendation: 0.0398 seconds
Masukkan judul atau deskripsi tugas akhir untuk rekomendasi:  TEKNIK PERBANDINGAN MACHINE LEARNING UNTUK PENYAKIT DIAGNOSIS DAN PREDIKSI JANTUNG
Rekomendasi berdasarkan input Anda (threshold=0.3):
- Title: PERBANDINGAN TEKNIK MACHINE LEARNING UNTUK DIAGNOSIS DAN PREDIKSI PENYAKIT JANTUNG
  Link: https://eprints.amikom.ac.id//id/eprint/26265/
  Creator: Akbar, Muhammad Ilham
  Date: 2024-03-19
  Similarity Score: 1.0000

- Title: ANALISIS KINERJA ALGORITMA SUPPORT VECTOR MACHINE DAN GRADIENT BOOSTING MACHINE UNTUK PREDIKSI PENYAKIT JANTUNG
  Link: https://eprints.amikom.ac.id//id/eprint/28815/
  Creator: Wirawan, Tegar
  Date: 2025-01-23
  Similarity Score: 0.4801

- Title: SISTEM PAKAR DIAGNOSA PENYAKIT JANTUNG MENGGUNAKAN METODE BAYES BERBASIS ANDROID
  Link: https://eprints.amikom.ac.id//id/eprint/9737/
  Creator: Faraditya, Riang Girinda
  Date: 2015-12-01
  Similarity Score: 0.4467

- Title: PERBANDINGAN MODEL ANALISIS SENTIMEN MENGGUNAKAN PENDEKATAN MACHINE LEARNING DAN DEEP LEARNING
  Link: https://eprints.amikom.ac.id//id/eprint/18313/
  Creator: Ikhsan, Muhammad
  Date: 2023-01-27
  Similarity Score: 0.4374

- Title: IMPLEMENTASI LOGISTIC REGRESSION UNTUK PREDIKSI PENYAKIT JANTUNG PADA DATASET CLEVELAND UCI
  Link: https://eprints.amikom.ac.id//id/eprint/28829/
  Creator: Al-Qodiry, Sayidah Aisah
  Date: 2025-01-21
  Similarity Score: 0.4321
```

Gambar 4 Hasil skenario 2

C. Akuisisi dan *Preprocessing* Data

Penelitian ini menggunakan metadata tugas akhir mahasiswa yang diperoleh melalui protokol OAI-PMH, yakni standar terbuka untuk pengambilan *metadata* dari berbagai repositori digital secara otomatis. Metadata tersebut mencakup judul, abstrak, dan kata kunci yang kemudian diproses dengan tahapan *preprocessing* seperti konversi ke huruf kecil, penghapusan tanda baca, angka, dan *stopword*, serta normalisasi kata untuk menghasilkan teks yang bersih dan konsisten. Ekstraksi fitur dilakukan dengan metode TF-IDF untuk representasi numerik teks sesuai dengan penelitian terkait sistem rekomendasi berbasis konten [8].

D. Perhitungan Kemiripan dengan *Cosine Similarity*

Kemiripan antar dokumen dihitung menggunakan metode *cosine similarity* yang mengukur sudut antar vektor TF-IDF dalam ruang multidimensi. Nilai *cosine similarity* yang diperoleh berada pada rentang -1 hingga 1 dan menjadi dasar rekomendasi tugas akhir dalam sistem [3], [5].

E. Evaluasi Sistem Rekomendasi

Evaluasi sistem dilakukan dengan mengukur *precision* dan *recall* menggunakan *confusion matrix* yang divisualisasikan sebagai *heatmap*. Hasil menunjukkan bahwa nilai-nilai pada diagonal utama (*True Positive* dan *True Negative*) mendominasi, mencerminkan efektivitas sistem dalam mengenali relevansi topik rekomendasi [16]. Gambar 5 menunjukkan hasil *confusion matrix*.



Gambar 5 Hasil *confusion matrix*

Pada eksperimen ini, sistem rekomendasi dievaluasi berdasarkan kemampuannya dalam menghasilkan rekomendasi tugas akhir yang relevan menggunakan pendekatan *Content-Based Filtering* dengan *Cosine Similarity*. Evaluasi dilakukan melalui visualisasi *confusion matrix* (*heatmap*), distribusi waktu respon, dan distribusi skor *similarity*.

Confusion matrix menggambarkan perbandingan antara label aktual dan prediksi sistem, yang mencakup empat kategori evaluasi:

1. *True Positive* (TP): Rekomendasi relevan yang diprediksi dengan benar, menunjukkan akurasi sistem.
2. *True Negative* (TN): Sistem berhasil tidak merekomendasikan item yang tidak relevan.
3. *False Positive* (FP): Item yang diprediksi relevan tetapi sebenarnya tidak, nilai rendah pada kategori ini menunjukkan sistem minim kesalahan.
4. *False Negative* (FN): Item relevan yang tidak direkomendasikan, nilai rendah menunjukkan sensitivitas sistem yang baik.

Confusion matrix atau dalam bahasa Indonesia disebut matriks kebingungan merupakan alat evaluasi yang penting dalam mengukur kinerja sistem klasifikasi seperti sistem rekomendasi. Matriks ini menyajikan data dalam format tabular yang membandingkan antara hasil prediksi sistem dan label aktual dari data uji. Dalam konteks sistem rekomendasi tugas akhir, *confusion matrix* digunakan untuk mengetahui seberapa banyak topik yang direkomendasikan oleh sistem sesuai atau tidak sesuai dengan relevansi sebenarnya.

Dalam visualisasi *heatmap* pada Gambar 5, nilai pada sumbu diagonal utama (yang mencerminkan nilai TP dan TN) memiliki intensitas warna yang lebih tinggi dibandingkan sel-sel lainnya. Hal ini menunjukkan bahwa sistem secara

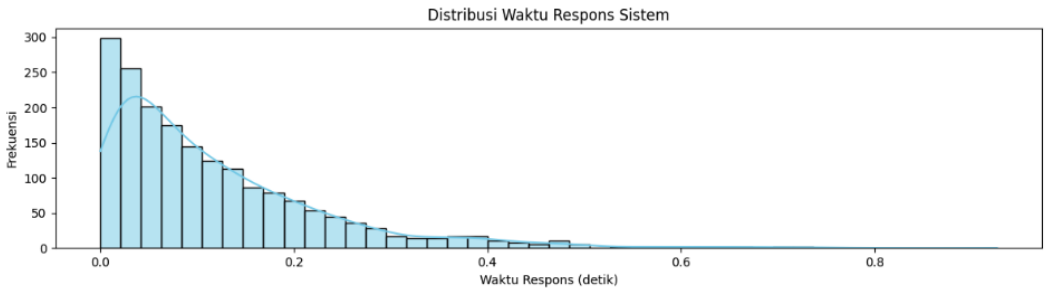
konsisten mampu memberikan rekomendasi yang benar (relevan), sekaligus menghindari memberikan rekomendasi yang salah (tidak relevan). Dominasi nilai tinggi pada TP dan TN menunjukkan bahwa sistem mampu membedakan antara rekomendasi yang relevan dan tidak relevan secara efektif, serta memiliki potensi tinggi untuk digunakan dalam lingkungan akademik nyata. Hal ini sejalan dengan temuan pada sistem rekomendasi berbasis *cosine similarity* yang efektif dalam meningkatkan akurasi [17]. Berdasarkan matriks kebingungan tersebut, metrik *precision* dan *recall* dihitung untuk mengukur kualitas rekomendasi:

1. *Precision* mengukur proporsi rekomendasi yang benar-benar relevan dari semua rekomendasi yang diberikan. Nilai *precision* yang tinggi menunjukkan sistem minim memberikan rekomendasi yang salah.
2. *Recall* mengukur proporsi rekomendasi relevan yang berhasil ditemukan oleh sistem dari seluruh rekomendasi relevan yang ada. Nilai *recall* yang tinggi menunjukkan sistem mampu menangkap sebagian besar rekomendasi yang relevan.

Hasil eksperimen menunjukkan nilai *precision* dan *recall* yang memuaskan, menandakan sistem rekomendasi efektif dalam memberikan rekomendasi yang sesuai dengan kebutuhan pengguna.

F. Distribusi Waktu Respon

Gambar 6 memperlihatkan histogram distribusi waktu respon sistem dalam menghasilkan rekomendasi. Sebagian besar proses rekomendasi berlangsung sangat cepat, dengan mayoritas waktu respon berada di bawah 0,2 detik. Pola distribusi menunjukkan tren menurun seiring bertambahnya waktu, yang mengindikasikan bahwa keterlambatan merupakan kejadian yang jarang terjadi. Kecepatan ini menjadi faktor penting dalam menjaga pengalaman pengguna, terutama saat sistem memproses sejumlah besar permintaan. Konsistensi performa dalam waktu respon yang rendah juga menunjukkan bahwa sistem tidak hanya akurat, tetapi juga efisien dan stabil, sehingga cocok untuk diterapkan di lingkungan produksi nyata.



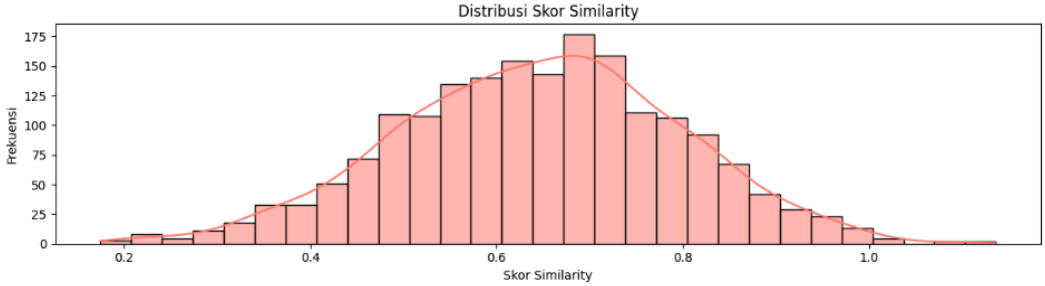
Gambar 6 Hasil *response time*

Histogram waktu respon memperlihatkan distribusi waktu yang dibutuhkan sistem untuk menghasilkan rekomendasi. Mayoritas waktu respon berada di bawah 0,2 detik yang menunjukkan performa sistem yang cepat dan efisien. Kecepatan ini sangat penting untuk menjaga pengalaman pengguna yang optimal, terutama saat

sistem harus menangani volume data yang besar. Studi terkait kinerja sistem rekomendasi *real-time* menunjukkan bahwa penerapan metode *cosine similarity* dan *collaborative filtering* dalam arsitektur *edge computing* dapat secara signifikan mengurangi latensi serta mempercepat waktu respon, jika dibandingkan dengan pendekatan *baseline* lainnya [18].

G. Distribusi Skor *Similarity*

Gambar 7 menunjukkan histogram distribusi skor *cosine similarity* yang menyerupai kurva normal, dengan konsentrasi skor pada rentang 0,4 hingga 0,8 dan puncaknya di sekitar 0,6. Pola ini mencerminkan bahwa sistem rekomendasi menghasilkan nilai kemiripan yang stabil dan konsisten antar dokumen. Rentang skor yang moderat hingga tinggi menandakan kemampuan sistem dalam membedakan tingkat kemiripan konten secara efektif, sehingga hasil rekomendasi tetap relevan dan tidak terdistorsi oleh nilai ekstrim. Distribusi ini memperkuat efektivitas pendekatan *Content-Based Filtering* dengan *Cosine Similarity* dalam mengenali kesamaan konten tugas akhir mahasiswa secara kontekstual dan bermakna.



Gambar 7 Hasil *similarity*

Tabel 1 Hasil dan perbandingan kinerja

Aspek Evaluasi	Dataset 2024 (1849 data)	Seluruh Data (28803 data)
Average Precision@5	0,6004	0,8827
Average Recall@5	0,5394	0,5022
Average Response Time (s)	0,0021	0,0398

Dengan mengacu pada Gambar 3, Gambar 4, dan Tabel 1 hasil dari kedua skenario di atas, skenario pertama menunjukkan bahwa sistem menghasilkan nilai *Average Precision*@5 sebesar 0,6004 dan *Recall*@5 sebesar 0,5394, dengan rata-rata waktu rekomendasi 0,0021 detik. Sementara itu, pada skenario kedua, *precision* meningkat signifikan menjadi 0,8827, meskipun *recall* sedikit menurun menjadi 0,5022, dengan waktu eksekusi meningkat ke 0,0398 detik per rekomendasi.

Peningkatan *precision* menunjukkan bahwa sistem lebih mampu memilih rekomendasi yang benar-benar relevan ketika jumlah data lebih besar. Namun, penurunan *recall* dapat terjadi karena sistem hanya memilih dokumen dengan skor kemiripan tertinggi, meninggalkan dokumen relevan lainnya di luar top-5 hasil

rekomendasi. Kenaikan waktu eksekusi merupakan *trade-off* yang wajar mengingat skala data yang jauh lebih besar, meskipun sistem tetap dapat merespon dalam waktu di bawah 0,05 detik, yang masih tergolong sangat cepat dan efisien.

Temuan ini sejalan dengan hasil penelitian oleh Zhang dan Zhao, yang menyatakan bahwa *cosine similarity* efektif dalam meningkatkan *precision* pada sistem rekomendasi berbasis dokumen akademik [19]. Selain itu, studi oleh Ahmed et al. juga menegaskan bahwa representasi teks dengan TF-IDF tetap menjadi pilihan unggul dalam klasifikasi dan rekomendasi berbasis konten karena efisiensi dan akurasinya [20]. Dari sisi waktu proses, performa sistem pada skenario kedua sejalan dengan penelitian oleh Gul-E-Laraib et al., yang menunjukkan bahwa pendekatan *cosine similarity* menghasilkan latensi rendah dalam sistem *edge computing* untuk personalisasi konten secara *real-time* [18].

Secara keseluruhan, hasil pengujian memperkuat bahwa metode TF-IDF dengan *cosine similarity* yang diterapkan pada sistem ini bersifat efektif dan efisien. Distribusi skor *cosine similarity* yang menyerupai kurva normal, dengan konsentrasi skor pada rentang 0,4 hingga 0,8 dan puncak sekitar 0,6, menunjukkan bahwa sistem mampu membedakan kemiripan konten antar dokumen dengan konsisten dan stabil. Ini memperkuat efektivitas pendekatan *Content-Based Filtering* berbasis *cosine similarity* dalam mengenali hubungan semantik antar topik tugas akhir mahasiswa [19]. Selain itu, sistem menunjukkan rata-rata waktu respon sebesar 0,0054 detik, yang mencerminkan efisiensi tinggi dan kelayakan untuk diimplementasikan dalam skenario layanan rekomendasi berbasis *web* atau aplikasi interaktif [18].

Namun demikian, penelitian ini juga memiliki beberapa keterbatasan yang perlu dicatat. Salah satunya adalah penurunan *recall* saat sistem diterapkan pada dataset yang lebih besar. Meskipun *precision* meningkat, *recall* sedikit menurun, yang mengindikasikan bahwa sistem cenderung hanya memilih dokumen dengan kemiripan tertinggi dan berpotensi melewatkan dokumen lain yang juga relevan tetapi berada di bawah peringkat atas. Hal ini merupakan konsekuensi umum dari pendekatan *Content-Based Filtering* yang mengandalkan *threshold* kemiripan tertentu.

Keterbatasan lain terletak pada ketergantungan sistem terhadap representasi berbasis TF-IDF, yang meskipun efisien, tidak mampu menangkap makna kontekstual secara mendalam. Ini dapat menjadi hambatan ketika sistem harus menangani teks dengan sinonim atau makna ganda. Selain itu, karena sistem ini murni berbasis konten, maka tidak mempertimbangkan preferensi eksplisit pengguna, seperti riwayat klik atau *rating*, yang dapat menjadi informasi tambahan penting dalam sistem rekomendasi modern. Terakhir, evaluasi sistem belum melibatkan pengguna akhir secara langsung, sehingga aspek subjektif seperti kepuasan pengguna, kemudahan pemahaman, atau relevansi perseptual belum diuji secara menyeluruh.

Simpulan

Penelitian ini berhasil mengembangkan sistem rekomendasi tugas akhir mahasiswa berbasis *Content-Based Filtering* dengan pendekatan *Cosine Similarity*. Sistem ini mampu memberikan rekomendasi yang relevan dan akurat dengan waktu respon yang sangat cepat, serta menunjukkan performa *precision* dan *recall* yang memadai pada berbagai skenario data. Distribusi skor *similarity* yang stabil serta ketahanan terhadap variasi *input* menunjukkan bahwa sistem ini konsisten dan fleksibel dalam mengenali kemiripan konten tugas akhir. Keunggulan sistem terletak pada efisiensi pemrosesan, akurasi rekomendasi, serta integrasi otomatis dengan data dari repositori institusi menggunakan protokol OAI-PMH. Meskipun demikian, sistem masih memiliki keterbatasan dalam hal *recall* dan belum memanfaatkan preferensi pengguna secara eksplisit.

Ke depan, sistem ini dapat dikembangkan lebih lanjut melalui integrasi pendekatan *hybrid filtering*, peningkatan teknik representasi teks menggunakan model NLP modern seperti BERT, serta evaluasi lebih komprehensif dengan melibatkan pengguna langsung. Dengan demikian, sistem ini memberikan kontribusi nyata terhadap pengembangan solusi rekomendasi akademik yang efisien, adaptif, dan dapat diimplementasikan secara luas di institusi pendidikan tinggi.

Daftar Pustaka

[1] R. V. K. Bevara *et al.*, “Resume2Vec: Transforming Applicant Tracking Systems with Intelligent Resume Embeddings for Precise Candidate Matching,” *Electronics (Switzerland)*, vol. 14, no. 4, Feb. 2025, doi: 10.3390/electronics14040794.

[2] D. El Alaoui, J. Riffi, A. Sabri, B. Aghoutane, A. Yahyaouy, and H. Tairi, “Comparative Study of Filtering Methods for Scientific Research Article Recommendations,” *Big Data and Cognitive Computing*, vol. 8, no. 12, Dec. 2024, doi: 10.3390/bdcc8120190.

[3] Q. Chen and O. Zhang, “Construction and Study of Textual Association Network Based on Cosine Similarity Algorithm,” in *Proceedings - 2023 3rd International Signal Processing, Communications and Engineering Management Conference, ISPCEM 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 829–835. doi: 10.1109/ISPCEM60569.2023.00156.

[4] Z. Fitri, Safriana, and Nurdin, “Computational thinking-based informatics material recommendation system for vocational school students with the content-based filtering method,” *Edelweiss Applied Science and Technology*, vol. 8, no. 6, pp. 4765–4777, 2024, doi: 10.55214/25768484.v8i6.3028.

[5] H. Yao, Y. Huang, J. Hu, and W. Xie, “Cosine similarity distance pruning algorithm Based on graph attention mechanism,” in *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, Institute of

Electrical and Electronics Engineers Inc., Dec. 2020, pp. 3311–3318. doi: 10.1109/BigData50022.2020.9378189.

[6] Y. Gao, H. Zhao, Q. Zhou, M. Qiu, and M. Liu, "An Improved News Recommendation Algorithm Based on Text Similarity," in *Proceedings - 2020 3rd International Conference on Smart BlockChain, SmartBlock 2020*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 132–136. doi: 10.1109/SmartBlock52591.2020.00031.

[7] K. E. Permana, "Comparison of User Based and Item Based Collaborative Filtering in Restaurant Recommendation System," *Mathematical Modelling of Engineering Problems*, vol. 11, no. 7, pp. 1922–1928, Jul. 2024, doi: 10.18280/mmep.110723.

[8] E. S. Negara, Sulaiman, R. Andryani, P. H. Saksono, and Y. Widyanti, "Recommendation System with Content-Based Filtering in NFT Marketplace," *Journal of Advances in Information Technology*, vol. 14, no. 3, pp. 518–522, 2023, doi: 10.12720/jait.14.3.518-522.

[9] S. Salloum and D. Rajamanthri, "Implementation and evaluation of movie recommender systems using collaborative filtering," *Journal of Advances in Information Technology*, vol. 12, no. 3, pp. 189–196, Aug. 2021, doi: 10.12720/jait.12.3.189-196.

[10] J. P. Yu and Y. Li, "Hybrid Recommendation Algorithm Combining User Similarity And Item Similarity," in *Proceedings - 2022 4th International Conference on Machine Learning, Big Data and Business Intelligence, MLBDBI 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 258–263. doi: 10.1109/MLBDBI58171.2022.00057.

[11] T. C. T. Tran, L. P. Phan, and H. X. Huynh, "Approach of Item-Based Collaborative Filtering Recommendation Using Energy Distance," *Journal of Advances in Information Technology*, vol. 15, no. 1, pp. 10–16, 2024, doi: 10.12720/jait.15.1.10-16.

[12] R. Widayanti, M. Heru, R. Chakim, C. Lukita, U. Rahardja, and N. Lutfiani, "Improving Recommender Systems using Hybrid Techniques of Collaborative Filtering and Content-Based Filtering," *Journal of Applied Data Sciences*, vol. 4, no. 3, pp. 289–302, 2023.

[13] W. B. Zulfikar, M. Irfan, M. Ghufron, Jumadi, and E. Firmansyah, "Marketplace affiliates potential analysis using cosine similarity and vision-based page segmentation," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 6, pp. 2492–2498, Dec. 2020, doi: 10.11591/eei.v9i6.2018.

[14] A. Feng, "Automatic Density Peaks Clustering based on the Cosine Similarity," in *Proceedings - 2022 4th International Conference on Artificial Intelligence and Advanced Manufacturing, AIAM 2022*, Institute of

Electrical and Electronics Engineers Inc., 2022, pp. 411–417. doi: 10.1109/AIAM57466.2022.00084.

[15] W. Lei and Z. Meng, "Text similarity calculation method of Siamese network based on ALBERT," in *Proceedings - 2022 International Conference on Machine Learning and Knowledge Engineering, MLKE 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 251–255. doi: 10.1109/MLKE55170.2022.00055.

[16] D. Meng and L. Wang, "Improving BERTScore via Word Similarity Matrix Iterative Alignment Method," in *2023 8th International Conference on Intelligent Computing and Signal Processing, ICSP 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 220–224. doi: 10.1109/ICSP58490.2023.10248840.

[17] M. N. Noori, J. Ahamed, and M. Ahmed, "Matrix Factorization and Cosine Similarity Based Recommendation System For Cold Start Problem in E-Commerce Industries," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 775–787, 2024, doi: 10.12785/ijcds/150156.

[18] Gul-E-Laraib *et al.*, "Content Caching in Mobile Edge Computing Based on User Location and Preferences Using Cosine Similarity and Collaborative Filtering," *Electronics (Switzerland)*, vol. 12, no. 2, Jan. 2023, doi: 10.3390/electronics12020284.

[19] Y. Zhang and H. Zhao, "Improving Precision in Recommendation with Cosine-Based Similarity on Educational Data," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 7, pp. 65–74, 2021, doi: 10.3991/ijet.v16i07.20959.

[20] N. Ahmed, R. Pathan, and M. Uddin, "Performance of TF-IDF Algorithm in Text Classification," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021, doi: 10.14569/IJACSA.2021.0120564.