Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 5 No. 1 (2026) pp: 3400-3408
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Perbandingan Algoritma K-Nearest Neighbor dan Cosine Similarity Untuk Sistem Rekomendasi Kursus Udemy

Ayudha Kusuma Ramadhani
Teknik Informatika, STMIK Widya Cipta Dharma
ayudha.kusuma.rahmadhani@gmail.com

Pertumbuhan pesat platform pembelajaran daring mendorong meningkatnya kebutuhan akan sistem rekomendasi yang mampu menyajikan kursus secara relevan dan personal. Udemy sebagai salah satu platform e-learning terbesar menyediakan ribuan kursus dengan variasi topik yang luas, sehingga pengguna sering mengalami kesulitan dalam menemukan kursus yang sesuai dengan minat dan kebutuhannya. Penelitian ini bertujuan untuk membandingkan kinerja algoritma Cosine Similarity dan K-Nearest Neighbor (KNN) dalam sistem rekomendasi kursus berbasis konten pada platform Udemy. Dataset yang digunakan terdiri dari 3.678 data kursus yang diperoleh dari Kaggle, dengan atribut judul kursus sebagai fitur utama. Data teks dipraproses dan direpresentasikan menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Cosine Similarity digunakan untuk mengukur tingkat kemiripan antar kursus, sedangkan KNN berfungsi sebagai mekanisme pencarian tetangga terdekat berdasarkan skor kemiripan. Evaluasi kinerja sistem rekomendasi dilakukan menggunakan metrik Precision@N dengan acuan kesesuaian kategori kursus. Hasil penelitian menunjukkan bahwa algoritma KNN menghasilkan tingkat relevansi rekomendasi yang lebih tinggi dibandingkan Cosine Similarity, namun membutuhkan waktu komputasi yang lebih besar. Sebaliknya, Cosine Similarity unggul dari sisi efisiensi dan stabilitas hasil rekomendasi. Temuan ini menunjukkan adanya trade-off antara relevansi dan efisiensi komputasi, sehingga pemilihan metode sistem rekomendasi perlu disesuaikan dengan kebutuhan dan skala implementasi. Penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem rekomendasi kursus daring berbasis konten.

Kata kunci: *K-Nearest Neighbor, Cosine Similarity, Sistem Rekomendasi, Udemy, Kursus.*

1. Latar Belakang

Perkembangan kursus daring menjadikan *platform* pembelajaran seperti Udemy menyediakan jumlah kursus yang sangat beragam, sehingga pengguna sering mengalami kesulitan dalam menemukan kursus yang sesuai dengan kebutuhan dan minat mereka. Kondisi ini mendorong perlunya sistem rekomendasi yang efektif untuk membantu proses pencarian dan pengambilan keputusan. Sistem rekomendasi berbasis konten berfungsi menyajikan item yang relevan berdasarkan karakteristik konten tanpa bergantung pada data interaksi pengguna, sehingga sesuai diterapkan pada *platform* pembelajaran daring dengan keterbatasan data pengguna [1]-[2]. Selain itu, perkembangan teknologi kecerdasan buatan pada lingkungan e-learning turut mendorong pemanfaatan berbagai pendekatan rekomendasi yang semakin adaptif dan personal [3].

Pemilihan algoritma rekomendasi menjadi faktor penting yang mempengaruhi kualitas hasil rekomendasi. Algoritma *Cosine Similarity* dan *K-Nearest Neighbor* (KNN) merupakan dua pendekatan yang umum digunakan untuk mengukur kemiripan antar item berbasis teks. *Cosine Similarity* menghitung tingkat kesamaan berdasarkan sudut antar vektor representasi data, sedangkan KNN berfungsi sebagai mekanisme pencarian tetangga terdekat berdasarkan nilai kemiripan tertentu [4]-[5]. Penerapan algoritma KNN telah menunjukkan kinerja yang baik pada berbagai domain rekomendasi lain, seperti sistem rekomendasi buku pada *platform* ritel, yang membuktikan efektivitas pendekatan *nearest neighbor* dalam mengidentifikasi item dengan karakteristik serupa [6].

Berbagai penelitian menunjukkan bahwa pendekatan *content-based filtering* berbasis TF-IDF dan *Cosine Similarity* telah berhasil diterapkan pada *domain* selain kursus daring, seperti rekomendasi resep masakan, artikel ilmiah, dan berita [7]-[9]. Meskipun demikian, kajian yang secara khusus membandingkan kinerja *Cosine Similarity* dan KNN dalam konteks sistem rekomendasi kursus daring berbasis konten masih terbatas. Hal ini menjadi tantangan tersendiri mengingat karakteristik data kursus daring umumnya didominasi oleh informasi teks singkat, seperti judul kursus, yang memerlukan evaluasi khusus terhadap efektivitas metode pengukuran kemiripan [10].

Berdasarkan kondisi tersebut, *research gap* penelitian ini terletak pada belum banyaknya kajian komparatif yang mengevaluasi performa algoritma *K-Nearest Neighbor* dan *Cosine Similarity* secara langsung dalam sistem rekomendasi kursus daring berbasis konten menggunakan dataset kursus Udemy. Oleh karena itu, rumusan masalah dalam penelitian yakni "Bagaimana perbandingan antara algoritma *K-Nearest Neighbor* dan *Cosine Similarity* dalam konteks sistem rekomendasi kursus di Udemy?". Sejalan dengan rumusan masalah tersebut, penelitian ini bertujuan untuk membandingkan performa kedua algoritma tersebut guna mengidentifikasi kelebihan dan keterbatasannya sebagai dasar pemilihan metode yang tepat dalam meningkatkan relevansi dan efektivitas rekomendasi kursus pada *platform* pembelajaran daring.

Secara lebih detail, terdapat beberapa batasan masalah yang diterapkan dalam penelitian ini sebagai berikut:

a. Data yang digunakan dalam penelitian ini terkait dengan preferensi dan interaksi pengguna terhadap kursus di Udemy.
b. Fokus pada perbandingan performa antara algoritma *K-Nearest Neighbor* dan *Cosine Similarity* dalam menghasilkan rekomendasi kursus yang akurat.
c. Pengujian hanya mencakup akurasi dan ketepatan eksekusi algoritma.

2. Metode Penelitian

2.1. Dataset dan Sumber

Data yang digunakan dalam penelitian ini merupakan data kursus daring yang diperoleh dari *platform* Kaggle melalui tautan https://www.kaggle.com/datasets/andrewmvd/udemy-courses. Dataset tersebut berisi 3.678 data kursus dengan 12 atribut yang mencakup informasi identitas kursus, judul kursus, kategori, tingkat kursus, harga, jumlah peserta, jumlah ulasan, serta durasi konten. Seluruh data tidak memiliki nilai kosong, sehingga tidak diperlukan proses imputasi data pada tahap prapemrosesan.

Tabel 1. Atribut Dataset Kursus Udemy

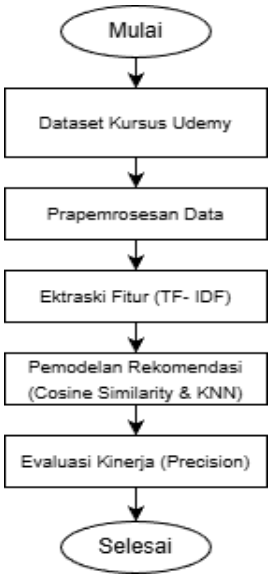
No	Atribut	Tipe Data	Digunakan	Keterangan
1	course_id	Integer	Tidak	Identitas unik kursus
2	course_title	Teks	Ya	Fitur utama untuk perhitungan kemiripan berbasis TF-IDF
3	url	Teks	Tidak	Tautan kursus
4	is_paid	Boolean	Tidak	Status berbayar atau gratis
5	price	Numerik	Tidak	Harga kursus
6	num_subscribers	Numerik	Tidak	Jumlah peserta kursus
7	num_reviews	Numerik	Tidak	Jumlah ulasan kursus
8	num_lectures	Numerik	Tidak	Jumlah materi kursus
9	level	Kategori	Tidak	Tingkat kesulitan kursus
10	content_duration	Numerik	Tidak	Durasi total kursus (jam)
11	published_timestamp	Teks	Tidak	Waktu publikasi kursus
12	subject	Kategori	Ya	Digunakan sebagai dasar evaluasi relevansi rekomendasi

Berdasarkan Tabel 1, atribut course_title digunakan sebagai fitur utama dalam pembentukan representasi teks menggunakan metode TF-IDF karena merepresentasikan karakteristik konten kursus secara ringkas dan informatif. Pendekatan TF-IDF telah banyak diterapkan dalam sistem rekomendasi berbasis teks untuk meningkatkan kualitas pengukuran kemiripan antar item [9]-[10]. Sementara itu, atribut subject dimanfaatkan sebagai acuan dalam proses evaluasi relevansi rekomendasi. Atribut lainnya tidak digunakan secara langsung dalam perhitungan kemiripan, namun tetap dipertahankan sebagai informasi deskriptif untuk memberikan gambaran umum mengenai karakteristik *dataset*.

2.2. Tahap Penelitian

Penelitian ini dilakukan melalui tahapan yang tersusun secara sistematis seperti yang terlihat pada Gambar 1. Tahap awal dimulai dengan pengumpulan dataset kursus Udemy yang diperoleh dari platform Kaggle. Selanjutnya dilakukan prapemrosesan data pada atribut judul kursus untuk membersihkan teks dan mengurangi noise. Data teks yang telah diproses kemudian direpresentasikan ke dalam bentuk vektor numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). Representasi ini digunakan untuk menghitung tingkat kemiripan antar kursus menggunakan metode *Cosine Similarity* dan algoritma *K-Nearest Neighbor* (KNN). Tahap akhir penelitian adalah evaluasi kinerja sistem rekomendasi menggunakan metrik precision untuk menilai relevansi hasil rekomendasi yang dihasilkan. Tahapan penelitian ini mengacu pada pendekatan sistem rekomendasi berbasis konten yang telah diterapkan dalam penelitian di Indonesia.

DOI: https://doi.org/10.31004/riggs.v5i1.6411



Gambar 1. *Work Flow* tahap penelitian

Keterangan: Alur tahapan dalam penelitian ini juga memastikan bahwa seluruh proses rekomendasi dilakukan secara konsisten berbasis konten tanpa melibatkan data historis pengguna.

2.3. Prapemrosesan Data

Prapemrosesan data dilakukan untuk menyiapkan data agar dapat direpresentasikan secara optimal dalam bentuk fitur numerik. *Dataset* kursus Udemy yang digunakan pada penelitian ini tidak memiliki nilai kosong pada setiap atribut, sehingga tidak diperlukan proses imputasi data. Meskipun demikian, tahap prapemrosesan tetap dilakukan untuk menyesuaikan data dengan kebutuhan analisis teks dan perhitungan kemiripan pada sistem rekomendasi berbasis konten.

Prapemrosesan difokuskan pada atribut course_title sebagai fitur utama sistem rekomendasi. Tahapan yang diterapkan meliputi case folding untuk menyeragamkan teks, penghapusan karakter khusus dan tanda baca, serta penghilangan *stopwords* yang tidak memberikan kontribusi semantik signifikan terhadap judul kursus. Proses ini bertujuan untuk mengurangi *noise* dan meningkatkan konsistensi representasi teks, sehingga kualitas vektor fitur yang dihasilkan menjadi lebih baik.

Hasil prapemrosesan berupa judul kursus yang telah dinormalisasi selanjutnya digunakan sebagai masukan pada tahap representasi fitur menggunakan metode Term *Frequency–Inverse Document Frequency* (TF-IDF). Pendekatan ini sejalan dengan penelitian sistem rekomendasi berbasis konten di Indonesia yang menekankan pentingnya pembersihan teks sebelum ekstraksi fitur untuk meningkatkan akurasi pengukuran kemiripan dan kualitas rekomendasi [9]-[10].

2.4. Representasi Data dengan TF-IDF

TF-IDF (*Term Frequency–Inverse Document Frequency*) merupakan metode pembobotan kata yang digunakan untuk merepresentasikan data teks ke dalam bentuk numerik. Metode ini memberikan bobot pada setiap kata berdasarkan frekuensi kemunculannya dalam suatu dokumen serta tingkat kepentingannya dalam keseluruhan kumpulan dokumen. Metode TF-IDF dan Cosine Similarity juga telah digunakan dalam berbagai domain klasifikasi teks, termasuk klasifikasi judul dokumen akademik [12]. Pendekatan ini menunjukkan efektivitas dalam mengukur kemiripan berbasis konten.

Dalam penelitian ini, TF-IDF digunakan untuk merepresentasikan judul kursus (course_title) sebagai dasar perhitungan kemiripan dalam sistem rekomendasi berbasis konten, sebagaimana umum diterapkan pada sistem rekomendasi berbasis teks [7], [10].

TF-IDF menggabungkan dua komponen utama, yaitu *Term Frequency* (TF) yang menunjukkan seberapa sering suatu kata muncul dalam dokumen, dan *Inverse Document Frequency* (IDF) yang mengukur tingkat kepentingan kata tersebut dalam keseluruhan dokumen. Nilai IDF dirumuskan sebagai berikut:

$$\text{IDF(w)} = log\ \frac{N}{DF(w)}$$

DOI: https://doi.org/10.31004/riggs.v5i1.6411

dengan N menyatakan jumlah seluruh dokumen dan DF(w) menyatakan jumlah dokumen yang mengandung kata w. Selanjutnya, nilai TF-IDF dihitung dengan mengalikan nilai TF dan IDF, yang dirumuskan sebagai berikut:

$$\text{TF-IDF (w, d)} = TF(w, d) \times IDF\ (w)$$

Hasil pembobotan TF-IDF menghasilkan vektor fitur yang merepresentasikan setiap judul kursus secara numerik. Vektor fitur ini kemudian digunakan sebagai masukan pada tahap perhitungan kemiripan menggunakan metode *Cosine Similarity* serta sebagai dasar pemilihan tetangga terdekat pada algoritma *K-Nearest Neighbor* (KNN). Implementasi TF-IDF pada penelitian ini dilakukan menggunakan *TfidfVectorizer* dari pustaka *scikit-learn* pada bahasa pemrograman *Python*. Pendekatan ini sejalan dengan berbagai penelitian sistem rekomendasi berbasis konten di Indonesia yang menunjukkan bahwa TF-IDF efektif dalam merepresentasikan teks singkat untuk mendukung proses pengukuran kemiripan dan rekomendasi item [8], [9], [11].

2.5. Algoritma *K-Nearest Neighbor* (KNN)

Algoritma *K-Nearest Neighbor* (KNN) digunakan dalam penelitian ini sebagai metode pembanding pada sistem rekomendasi berbasis konten. Berbeda dengan penerapan KNN pada permasalahan klasifikasi, KNN pada penelitian ini tidak digunakan untuk menentukan kelas tertentu, melainkan untuk memilih sejumlah k kursus terdekat berdasarkan tingkat kemiripan fitur. Pendekatan *content-based filtering* banyak digunakan pada berbagai domain karena mampu memberikan rekomendasi awal ketika data pengguna masih terbatas, khususnya pada kondisi *cold-start* [1].

Pada penelitian ini, setiap kursus direpresentasikan dalam bentuk vektor fitur hasil pembobotan TF-IDF terhadap atribut course_title. Representasi TF-IDF dipilih karena efektif dalam mengekstraksi karakteristik penting dari data teks dan umum digunakan dalam sistem rekomendasi berbasis konten, terutama pada data teks singkat seperti judul [7], [10]. Selanjutnya, tingkat kemiripan antara kursus acuan dan kursus lain dihitung menggunakan metode *Cosine Similarity*, yang mengukur kesamaan antar vektor fitur teks berdasarkan sudut vektor-nya [4], [13].

Perlu ditegaskan bahwa pada penelitian ini algoritma *K-Nearest Neighbor* tidak digunakan sebagai metode pembelajaran atau klasifikasi, melainkan sebagai mekanisme pencarian tetangga terdekat (*nearest neighbor retrieval*). KNN berfungsi untuk mengambil sejumlah kursus dengan nilai kemiripan tertinggi berdasarkan skor *similarity* yang dihasilkan dari representasi TF-IDF. Pendekatan ini sejalan dengan pemanfaatan KNN sebagai metode *instance-based* dalam berbagai sistem rekomendasi untuk menemukan item yang paling relevan berdasarkan kedekatan fitur, seperti pada sistem rekomendasi buku dan musik [5], [6], [14], sehingga rekomendasi yang dihasilkan tetap bersifat *content-based*.

2.6. *Cosine Similarity*

Cosine Similarity mengukur tingkat kemiripan antar dokumen berdasarkan sudut antar vektor representasi teks [13]. Metode ini banyak diterapkan pada data berbasis teks karena tidak dipengaruhi oleh panjang vektor, melainkan berfokus pada pola kemunculan fitur. Beberapa penelitian sebelumnya menunjukkan bahwa *Cosine Similarity* efektif dalam mengukur kesamaan data teks dan mampu menghasilkan kinerja yang baik pada sistem klasifikasi maupun sistem rekomendasi. Secara matematis, nilai *Cosine Similarity* antara dua vektor X dan Y dirumuskan sebagai berikut:

$$Cos(X, Y) = (\Sigma\ XiYi) / (\sqrt{(\Sigma\ Xi^2)} \times \sqrt{(\Sigma\ Yi^2)}$$

Nilai *Cosine Similarity* berada pada rentang 0 hingga 1, di mana nilai yang semakin mendekati 1 menunjukkan tingkat kemiripan data yang semakin tinggi. Penggunaan *Cosine Similarity* pada sistem rekomendasi berbasis konten terbukti memberikan hasil yang stabil dan efisien [12]. Oleh karena itu, metode ini sesuai digunakan untuk menentukan kedekatan antar data dalam sistem rekomendasi.

2.7. Metrik Evaluasi

Evaluasi kinerja sistem rekomendasi pada penelitian ini dilakukan untuk menilai tingkat relevansi rekomendasi kursus yang dihasilkan oleh metode *Cosine Similarity* dan *K-Nearest Neighbor* (KNN). Evaluasi difokuskan pada kesesuaian kategori (*subject*) antara kursus acuan dan kursus yang direkomendasikan, sebagaimana disajikan pada tabel hasil evaluasi. Pendekatan ini sesuai dengan karakteristik sistem rekomendasi berbasis konten yang mengandalkan kemiripan atribut item tanpa memanfaatkan riwayat interaksi pengguna [1].

DOI: https://doi.org/10.31004/riggs.v5i1.6411

Precision digunakan sebagai metrik evaluasi utama dan ditampilkan dalam bentuk nilai numerik pada tabel perbandingan performa metode. Suatu rekomendasi dinyatakan relevan apabila kursus yang direkomendasikan memiliki kategori yang sama dengan kursus acuan. Nilai *precision* dihitung berdasarkan proporsi kursus relevan pada hasil *Top-N recommendation* yang dihasilkan oleh masing-masing metode, sehingga memungkinkan perbandingan langsung antara *Cosine Similarity* dan KNN [10], [13].

Recall dan F1-Score digunakan sebagai indikator konseptual dalam pembahasan hasil untuk memberikan konteks terhadap nilai *precision* yang diperoleh. Kedua metrik tersebut tidak dihitung secara numerik karena penelitian ini tidak menggunakan *ground truth* eksplisit untuk setiap pengguna, namun tetap dimanfaatkan untuk menjelaskan karakteristik kelengkapan dan keseimbangan rekomendasi yang ditunjukkan pada bagian hasil [1].

Accuracy disajikan sebagai metrik tambahan untuk memberikan gambaran umum performa sistem, namun tidak dijadikan indikator utama dalam interpretasi hasil. Hal ini disebabkan karena sistem yang dikembangkan berfokus pada rekomendasi item, bukan klasifikasi kelas. Pendekatan evaluasi yang menekankan *precision* sebagai metrik utama ini sejalan dengan penelitian sistem rekomendasi berbasis *content-based filtering* di Indonesia yang lebih memprioritaskan relevansi rekomendasi dibandingkan ketepatan klasifikasi [15].

3. Hasil dan Diskusi

3.1. Hasil Prapemrosesan Data

Tahap prapemrosesan data dilakukan untuk menyiapkan data teks agar dapat digunakan secara optimal dalam sistem rekomendasi berbasis konten. Prapemrosesan difokuskan pada atribut course_title karena atribut ini merepresentasikan konten utama dari setiap kursus dan menjadi dasar dalam pembentukan fitur pada tahap selanjutnya.

Berdasarkan hasil prapemrosesan, seluruh judul kursus berhasil dinormalisasi melalui beberapa tahapan, yaitu pengubahan teks menjadi huruf kecil (*case folding*), penghapusan karakter khusus dan tanda baca, serta penghilangan kata-kata umum (*stopwords*) yang tidak memiliki kontribusi signifikan terhadap makna konten. Proses ini bertujuan untuk mengurangi noise dan menghasilkan representasi teks yang lebih konsisten.

Hasil prapemrosesan menunjukkan bahwa judul kursus yang sebelumnya memiliki variasi penulisan, angka, dan simbol telah disederhanakan menjadi kumpulan kata kunci yang lebih representatif terhadap topik kursus. Dengan demikian, tahap prapemrosesan ini berperan penting dalam meningkatkan kualitas data teks sebelum direpresentasikan ke dalam bentuk vektor menggunakan metode TF-IDF.Sebagai ilustrasi, Tabel 4.1 menampilkan contoh hasil prapemrosesan judul kursus sebelum dan sesudah dilakukan pembersihan teks.

Tabel 2. Contoh Hasil Prapemrosesan Judul Kursus

No	Judul Kursus Sebelum Prapemrosesan	Judul Kursus Sesudah Prapemrosesan
1	Learn Python Programming From Scratch!	learn python programming scratch
2	The Complete JavaScript Course 2021: From Zero to Expert	complete javascript course zero expert
3	Microsoft Excel – Advanced Excel Formulas & Functions	microsoft excel advanced excel formulas functions

3.2. Hasil Implementasi TF-IDF

Data teks hasil prapemrosesan pada atribut course_title direpresentasikan ke dalam bentuk vektor numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). Proses ini bertujuan untuk mengekstraksi kata-kata penting yang mampu merepresentasikan karakteristik setiap kursus secara lebih informatif. Implementasi TF-IDF menghasilkan sebuah matriks fitur berdimensi tinggi, di mana setiap baris merepresentasikan satu judul kursus dan setiap kolom merepresentasikan kata unik yang terbentuk setelah prapemrosesan. Bobot TF-IDF yang dihasilkan menunjukkan tingkat kepentingan suatu kata dalam membedakan satu kursus dengan kursus lainnya, sehingga kata-kata yang lebih spesifik terhadap topik kursus memiliki bobot yang lebih tinggi dibandingkan kata-kata umum. Selain itu, setiap kursus dipetakan ke dalam indeks tertentu untuk memudahkan proses pencarian dan pengambilan vektor fitur yang sesuai pada tahap perhitungan kemiripan. Representasi vektor TF-IDF ini menjadi dasar utama dalam mengukur tingkat kemiripan antar kursus menggunakan metode *Cosine Similarity* yang dibahas pada subbab berikutnya.

DOI: https://doi.org/10.31004/riggs.v5i1.6411

Tabel 3. Contoh Bobot TF-IDF pada Judul Kursus

Kata	Bobot TF-IDF
python	0.412
programming	0.356
beginner	0.298
course	0.121
advanced	0.087

**Keterangan: Bobot TF-IDF ditampilkan sebagai contoh untuk menunjukkan kontribusi relatif setiap kata dalam merepresentasikan judul kursus.

3.3. Hasil Rekomendasi Menggunakan *Cosine Similarity*

Pada tahap ini, sistem rekomendasi menghasilkan daftar kursus berdasarkan tingkat kemiripan judul kursus menggunakan metode *Cosine Similarity*. Perhitungan kemiripan dilakukan terhadap vektor TF-IDF yang diperoleh dari hasil representasi teks judul kursus. Setiap kursus direpresentasikan sebagai vektor numerik berdimensi tinggi, kemudian nilai *Cosine Similarity* dihitung untuk mengukur tingkat kesamaan antar kursus.

Berdasarkan hasil implementasi pada kode program, sistem menghitung nilai *Cosine Similarity* antara satu kursus acuan dengan seluruh kursus lainnya dalam dataset. Nilai kemiripan berada pada rentang 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan tingkat kemiripan konten yang lebih besar. Selanjutnya, kursus-kursus dengan nilai *Cosine Similarity* tertinggi dipilih sebagai rekomendasi (*Top-N recommendation*). Hasil pengujian menunjukkan bahwa kursus-kursus yang memiliki pola kata yang serupa pada judulnya cenderung memperoleh nilai kemiripan yang lebih tinggi. Sebagai contoh, untuk kursus dengan kata kunci "*easy steps*", sistem merekomendasikan kursus lain yang juga memiliki struktur judul dan kata dominan yang serupa.

Tabel 4.3 Hasil Rekomendasi Kursus Menggunakan Cosine Similarity

No	Judul Kursus Rekomendasi	Nilai Cosine Similarity
1	Easy Steps Photoshop Expert	0.355495
2	Kickstarter success 5 easy steps	0.294596
3	7 Steps Introduction Trading	0.272408
4	Setup MicroBlog 3 Easy Steps	0.268971
5	Guitar Master Guitar Fretboard 6 Easy Steps	0.263459
6	Learn Program JavaScript in easy steps	0.259641
7	Starting play blues harmonica easy steps	0.259312
8	Learn improvise harmonica nice easy steps	0.253829
9	Instant Harmonica single notes 5 easy steps	0.253557

Cosine Similarity digunakan untuk mengukur tingkat kemiripan antar judul kursus dengan membandingkan arah vektor fitur hasil pembobotan TF-IDF. Setiap judul kursus direpresentasikan sebagai vektor numerik yang mencerminkan bobot kepentingan kata-kata penyusunnya. Nilai *Cosine Similarity* dihitung berdasarkan perbandingan hasil perkalian titik dua vektor terhadap hasil perkalian panjang masing-masing vektor, sehingga menghasilkan nilai kemiripan dalam rentang 0 hingga 1. Semakin besar nilai *Cosine Similarity* yang diperoleh, semakin tinggi tingkat kesamaan konten antara judul kursus acuan dan kursus pembanding. Pada penelitian ini, nilai *Cosine Similarity* digunakan untuk mengurutkan kursus berdasarkan tingkat kemiripannya, di mana kursus dengan nilai tertinggi direkomendasikan sebagai kursus yang paling relevan dalam sistem rekomendasi berbasis konten.

3.4. Hasil Rekomendasi Menggunakan *K-Nearest Neighbor* (KKN)

Pada tahap ini, sistem rekomendasi diimplementasikan menggunakan algoritma *K-Nearest Neighbor* (KNN) untuk menentukan sejumlah kursus yang memiliki tingkat kemiripan tertinggi terhadap kursus yang dipilih pengguna. Algoritma KNN bekerja dengan mencari k tetangga terdekat berdasarkan representasi vektor TF-IDF dari judul kursus, sehingga rekomendasi yang dihasilkan bersifat content-based.

Dalam penelitian ini, ukuran kedekatan antar kursus dihitung menggunakan *cosine distance* sebagai metrik pada algoritma KNN. Nilai *cosine distance* kemudian dikonversi menjadi *cosine similarity* dengan rentang nilai antara 0 hingga 1, di mana nilai yang semakin mendekati 1 menunjukkan tingkat kemiripan konten yang semakin tinggi. Pendekatan ini memungkinkan hasil rekomendasi KNN untuk dibandingkan secara langsung dengan metode *Cosine Similarity* pada tahap analisis selanjutnya.

DOI: https://doi.org/10.31004/riggs.v5i1.6411

Berdasarkan hasil pengujian, sistem berhasil menghasilkan daftar rekomendasi kursus yang relevan dengan kursus acuan. Tabel 4.4 menampilkan sepuluh kursus dengan nilai kemiripan tertinggi yang direkomendasikan menggunakan algoritma KNN.

Tabel 4.3 Hasil Rekomendasi Kursus Menggunakan *KNN Similarity*

No	Judul Kursus Rekomendasi	Nilai KNN *Similarity*
1	Capital Market Immersion	0.8732
2	Buchführung Lernen Leicht Gemacht	0.8689
3	Flow Management Forecasting	0.8654
4	Curso Avanzado de Trading	0.8627
5	CFA Level Workshop Ethics Quantitative Methods	0.8591
6	8 Amortization Schedules	0.8568
7	Curso de Trading Práctico Basado en Casos Reales	0.8543
8	TeeterTotter Accounting	0.8526
9	Mortgage Acceleration	0.8502

**Keterangan: Tabel ini menunjukkan hasil rekomendasi kursus menggunakan algoritma *K-Nearest Neighbor* berdasarkan tingkat kemiripan konten judul kursus yang direpresentasikan dalam bentuk vektor TF-IDF. Nilai similarity score diperoleh dari konversi *cosine distance* menjadi *cosine similarity*.

3.5. Perbandingan Performa *Cosine Similarity* dan KNN

Perbandingan performa dilakukan untuk mengevaluasi kemampuan metode *Cosine Similarity* dan *K-Nearest Neighbor* (KNN) dalam menghasilkan rekomendasi kursus yang relevan. Kedua metode menggunakan representasi fitur yang sama, yaitu vektor TF-IDF dari judul kursus, sehingga perbedaan kinerja yang dihasilkan sepenuhnya dipengaruhi oleh mekanisme perhitungan kemiripan dan proses seleksi rekomendasi. *Cosine Similarity* menghitung tingkat kemiripan antara kursus acuan dan seluruh kursus lain secara langsung berdasarkan sudut antar vektor TF-IDF. Pendekatan ini bersifat global karena seluruh kursus dibandingkan tanpa proses pemilihan tetangga. Nilai kemiripan berada pada rentang 0 hingga 1, di mana nilai yang lebih besar menunjukkan tingkat kesamaan konten yang lebih tinggi. Pendekatan ini umum digunakan pada sistem rekomendasi berbasis teks karena efisiensi komputasinya yang tinggi [10]-[11].

Sebaliknya, algoritma *K-Nearest Neighbor* menentukan rekomendasi dengan memilih sejumlah k tetangga terdekat dari kursus acuan berdasarkan nilai kemiripan tertinggi. Pada penelitian ini, jarak antar vektor dihitung menggunakan *cosine distance*, kemudian dikonversi menjadi skor kesamaan. Proses ini memungkinkan KNN menghasilkan rekomendasi yang lebih terfokus pada lingkungan kursus dengan karakteristik konten yang serupa, sebagaimana dijelaskan dalam penelitian Bahrani et al. [5]. Evaluasi performa kedua metode dilakukan menggunakan metrik Precision@N. Precision@N mengukur proporsi rekomendasi yang relevan pada N rekomendasi teratas yang dihasilkan sistem. Suatu rekomendasi dianggap relevan apabila kursus yang direkomendasikan memiliki kategori (*subject*) yang sama dengan kursus acuan. Secara matematis, Precision@N dirumuskan sebagai berikut:

$$\text{Precision@N} = \frac{Jumlah\ rekomendasi\ relevan}{N}$$

Nilai Precision@N dihitung untuk setiap kursus uji, kemudian dirata-ratakan untuk memperoleh nilai performa akhir masing-masing metode.

Tabel 4. Perbandingan Performa Cosine Similarity dan KNN

Metode	Parameter	Precision@5	Precision@10	Waktu (detik)	Metode
Cosine	θ=0.5	0.3333	0.3333	0.0018	Cosine
Cosine	θ=0.6	0.3333	0.3333	0.0018	Cosine
Cosine	θ=0.7	0.3333	0.3333	0.0014	Cosine
KNN	k=3	1.0000	1.0000	0.0659	KNN
KNN	k=5	1.0000	1.0000	0.0698	KNN
KNN	k=7	1.0000	1.0000	0.0718	KNN

Berdasarkan hasil eksperimen yang dilakukan pada tiga judul kursus uji, algoritma K-Nearest Neighbor menunjukkan performa relevansi yang lebih tinggi dibandingkan *Cosine Similarity*. KNN dengan parameter k = 3 mencapai nilai Precision@5 dan Precision@10 sebesar 1,000, yang berarti seluruh rekomendasi teratas memiliki kategori (*subject*) yang sama dengan kursus acuan. Hasil serupa juga diperoleh pada variasi k = 5 dan k = 7, dengan

DOI: https://doi.org/10.31004/riggs.v5i1.6411

nilai Precision@10 tetap berada pada 1,000. Sebaliknya, metode *Cosine Similarity* menghasilkan nilai Precision@5 dan Precision@10 sebesar 0,333 pada seluruh variasi ambang kemiripan ($\theta = 0{,}5$; 0,6; dan 0,7). Nilai tersebut menunjukkan bahwa hanya sekitar sepertiga dari rekomendasi teratas yang memiliki kategori yang sama dengan kursus acuan. Dari sisi efisiensi komputasi, *Cosine Similarity* unggul dengan waktu eksekusi rata-rata sebesar 0,0016 detik per query, jauh lebih cepat dibandingkan KNN yang memerlukan waktu rata-rata 0,0692 detik per query. Perbedaan ini disebabkan oleh mekanisme KNN yang memerlukan proses pencarian tetangga terdekat secara eksplisit, sedangkan *Cosine Similarity* hanya melakukan perhitungan kemiripan vektor secara langsung.

Temuan ini menunjukkan adanya *trade-off* antara relevansi dan efisiensi komputasi. KNN lebih unggul dalam menghasilkan rekomendasi yang relevan, sedangkan *Cosine Similarity* lebih efisien untuk kebutuhan sistem dengan keterbatasan waktu respons. Oleh karena itu, pemilihan metode rekomendasi perlu disesuaikan dengan tujuan sistem, apakah lebih menekankan pada ketepatan rekomendasi atau kecepatan komputasi.

3.6. Analisis dan Interpretasi Hasil

Hasil penelitian menunjukkan bahwa representasi fitur teks menggunakan metode TF-IDF mampu merepresentasikan karakteristik utama judul kursus secara efektif, sehingga mendukung implementasi sistem rekomendasi berbasis konten. Proses prapemrosesan data, seperti pembersihan teks dan normalisasi, berperan dalam meningkatkan kualitas vektor fitur dan akurasi pengukuran kemiripan antar kursus.

Metode *Cosine Similarity* mampu mengidentifikasi kesamaan konten antar kursus secara langsung dengan efisiensi komputasi yang tinggi. Sementara itu, algoritma *K-Nearest Neighbor* (KNN) digunakan sebagai mekanisme pencarian tetangga terdekat berdasarkan skor kemiripan, sehingga menghasilkan rekomendasi yang lebih terfokus pada kelompok kursus dengan tingkat kemiripan tinggi. Hal ini tercermin dari nilai *precision* yang lebih tinggi pada KNN dibandingkan *Cosine Similarity*, sejalan dengan temuan pada penelitian sistem rekomendasi berbasis konten di berbagai domain [5], [13].

Hasil rekomendasi yang dihasilkan menunjukkan adanya rekomendasi lintas kategori (*subject*), yang disebabkan oleh keterbatasan fitur yang hanya menggunakan judul kursus sebagai representasi konten. Kondisi ini mengindikasikan adanya permasalahan *cold-start* yang umum terjadi pada sistem rekomendasi berbasis konten, khususnya ketika informasi item bersifat terbatas. Meskipun demikian, pendekatan ini tetap relevan untuk memberikan rekomendasi awal tanpa memerlukan data interaksi pengguna [1].

Gambar 2. Penerapan Aplikasi

Untuk memvalidasi hasil evaluasi secara praktis, metode yang diusulkan diimplementasikan dalam bentuk prototipe aplikasi berbasis *web* menggunakan *framework Streamlit*. Implementasi aplikasi tersebut ditunjukkan pada Gambar 2, yang menampilkan antarmuka pemilihan judul kursus acuan serta daftar rekomendasi kursus yang dihasilkan berdasarkan tingkat kemiripan. Hasil rekomendasi yang ditampilkan pada aplikasi konsisten dengan perhitungan dan evaluasi kuantitatif yang telah dibahas sebelumnya, sehingga menunjukkan bahwa metode yang diusulkan tidak hanya valid secara konseptual, tetapi juga dapat diterapkan pada skenario nyata sistem rekomendasi kursus daring.

4. Kesimpulan dan Saran

Penelitian ini mengkaji penerapan dan perbandingan algoritma *Cosine Similarity* dan *K-Nearest Neighbor* (KNN) dalam sistem rekomendasi kursus Udemy berbasis konten dengan pendekatan TF-IDF sebagai representasi fitur teks. Sistem rekomendasi dibangun dengan memanfaatkan informasi judul kursus dan atribut pendukung lainnya untuk mengukur tingkat kemiripan antar kursus. Hasil penelitian menunjukkan bahwa *Cosine Similarity* mampu

DOI: https://doi.org/10.31004/riggs.v5i1.6411

memberikan rekomendasi yang lebih stabil dan mudah diinterpretasikan karena secara langsung mengukur kesamaan arah vektor TF-IDF antar dokumen kursus. Nilai kemiripan yang dihasilkan bersifat kontinu sehingga memudahkan proses pemeringkatan rekomendasi berdasarkan tingkat relevansi. Sementara itu, K-Nearest Neighbor (KNN) menghasilkan rekomendasi berdasarkan kedekatan instance pada ruang fitur dan efektif dalam menemukan kursus yang memiliki karakteristik serupa. Namun, performa KNN sangat dipengaruhi oleh representasi data dan pemilihan parameter, sehingga hasil rekomendasi lebih bersifat relatif dan ditampilkan dalam bentuk peringkat tetangga terdekat. Berdasarkan hasil perbandingan, dapat disimpulkan bahwa Cosine Similarity lebih sesuai digunakan sebagai metode utama dalam sistem rekomendasi berbasis konten teks, sedangkan KNN berperan sebagai metode pembanding yang menekankan pendekatan *instance-based learning*. Kombinasi keduanya memberikan gambaran yang komprehensif mengenai perilaku sistem rekomendasi dalam konteks data kursus daring. Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran untuk pengembangan lebih lanjut. Penelitian selanjutnya disarankan untuk memperluas sumber fitur teks dengan menambahkan atribut seperti deskripsi kursus, *learning objectives*, atau *review* pengguna agar representasi TF-IDF menjadi lebih kaya dan informatif. Selain itu, metode rekomendasi dapat dikembangkan dengan mengintegrasikan pendekatan lain seperti word embedding atau model pembelajaran berbasis *neural network* untuk dibandingkan dengan *Cosine Similarity* dan KNN. Evaluasi sistem juga dapat diperluas menggunakan skenario *Top-K recommendation* dan uji pengguna (*user study*) agar hasil yang diperoleh lebih mencerminkan kebutuhan pengguna secara nyata. Pengembangan aplikasi sistem rekomendasi secara lebih lanjut, termasuk optimasi parameter dan peningkatan antarmuka pengguna, diharapkan dapat meningkatkan kegunaan sistem serta mendukung penerapannya pada skala data yang lebih besar.

Referensi

1. Z. Fayyaz, M. Ebrahimian, D. Nawara, A. Ibrahim, and R. Kashef, “Recommendation systems: Algorithms, challenges, metrics, and business opportunities,” Applied Sciences, vol. 10, no. 21, p. 7748, 2020, doi: 10.3390/app10217748.
2. Y. Christian and K. Kelvin, “Rancang bangun aplikasi kursus online berbasis web dengan sistem rekomendasi metode content-based filtering,” Rabit: Jurnal Teknologi dan Sistem Informasi Univrab, vol. 7, no. 1, 2022, doi: 10.36341/rabit.v7i1.2181.
3. L. El Youbi El Idrissi, I. Akharraz, and A. Ahaitouf, “Personalized e-learning recommender system based on autoencoders,” Applied System Innovation, vol. 6, no. 6, p. 102, 2023, doi: 10.3390/asi6060102.
4. R. Habibi, “Analisis sistem rekomendasi pekerjaan berbasis content-based filtering dengan cosine similarity,” Jurnal Informatika, 2022. [Online]. Available: https://ejurnal.ulbi.ac.id/index.php/informatika/article/view/2363
5. P. Bahrani, B. Minaei-Bidgoli, H. Parvin, M. Mirzarezaee, and A. Keshavarz, “A new improved KNN-based recommender system,” The Journal of Supercomputing, vol. 80, no. 1, pp. 800–834, 2024, doi: 10.1007/s11227-023-05447-1.
6. H. Dharmawan, T. Tukino, S. S. Hilabi, and I. Karniawulan, “Sistem rekomendasi buku dengan metode k-nearest neighbor (K-NN) pada Gramedia,” ZONAsi: Jurnal Sistem Informasi, vol. 5, no. 1, pp. 16–25, 2023, doi: 10.31849/zn.v5i1.12203.
7. N. N. K. Sari, R. Priskila, and P. B. A. A. Putra, “Implementasi content-based filtering menggunakan TF-IDF dan cosine similarity untuk sistem rekomendasi resep masakan,” Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika, 2024. [Online]. Available: https://jurnal.polgan.ac.id/index.php/sinkron/article/view/14778
8. A. Rianti, N. W. Abdul Majid, and A. Fauzi, “Machine learning journal article recommendation system using content-based filtering (TF-IDF & cosine similarity),” Jurnal Teknologi dan Sistem Informasi, vol. 22, no. 1, 2025, doi: 10.12962/j24068535.v22i1.a1193.
9. I. A. W. Nandita, “News recommendation system using content-based TF-IDF and cosine similarity,” Journal of Artificial Intelligence and Computing, 2025. [Online]. Available: https://jurnal.polibatam.ac.id/index.php/JAIC/article/view/9807
10. A. A. Huda, R. Fajarudin, and A. Hadinegoro, “Sistem rekomendasi content-based filtering menggunakan TF-IDF vector similarity untuk rekomendasi artikel berita,” Building of Informatics, Technology and Science (BITS), vol. 4, no. 3, pp. 1679–1686, 2022, doi: 10.47065/bits.v4i3.251.
11. M. Farhan, M. Rifky Andreawan, R. Antonius, and N. Dhiya Ulhag, “Evaluasi dan pengembangan sistem rekomendasi game berbasis content-based filtering dengan TF-IDF dan cosine similarity,” BINER: Jurnal Ilmu Komputer, Teknik dan Multimedia, vol. 2, no. 4, pp. 400–408, 2024.
12. Nurjanah, Husaini, and J. Salat, “Penggunaan metode cosine similarity dan TF-IDF untuk klasifikasi judul seminar proposal,” Sagita Academia Journal, vol. 2, no. 1, pp. 72–79, 2024, doi: 10.61579/sagita.v2i1.60.
13. F. Christyawan, A. Rohman, and A. Hartanto, “Application of content-based filtering method using cosine similarity in restaurant selection recommendation system,” Journal of Information Systems and Informatics, vol. 6, no. 3, pp. 1559–1576, 2024, doi: 10.51519/journalisi.v6i3.806.
14. M. A. Munajad, A. Ridwan, and T. G. Pratama, “Pengembangan sistem rekomendasi musik dengan K-Means dan K-Nearest Neighbors berbasis cosine similarity,” Sainteks Journal, vol. 22, no. 2, pp. 153–165, 2025, doi: 10.30595/sainteks.v22i2.27815.
15. S. Oyadila, D. Abdullah, and A. Razi, “Implementasi content-based filtering dengan TF-IDF dan cosine similarity untuk sistem rekomendasi destinasi wisata di Aceh Tengah,” Rabit: Jurnal Teknologi dan Sistem Informasi Univrab, vol. 10, no. 2, 2025, doi: 10.36341/rabit.v10i2.6532.

DOI: https://doi.org/10.31004/riggs.v5i1.6411