

# Sistem Rekomendasi Content-based Filtering Menggunakan TF-IDF Vector Similarity Untuk Rekomendasi Artikel Berita

Arif Akbarul Huda<sup>1,\*</sup>, Rohmad Fajarudin<sup>2</sup>, Arifiyanto Hadinegoro<sup>3</sup>

Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia

Email: <sup>1,\*</sup>arif.akbarul@amikom.ac.id, <sup>2</sup>rohmad.f@students.amikom.ac.id, <sup>3</sup>arifiyanto@amikom.ac.id

Email Penulis Korespondensi: arif.akbarul@amikom.ac.id

Submitted: 09/11/2022; Accepted: 30/12/2022; Published: 30/12/2022

**Abstrak**—Populasi mahasiswa aktif prodi Informatika Universitas Amikom Yogyakarta pada semester Ganjil 2021 sebanyak 3870 orang. Upaya pelacakan minat terhadap tiga pilihan konsentrasi dilakukan sejak dini, melalui rekomendasi literasi artikel. Ragam artikel diproduksi terus menerus dan diberikan secara berkelanjutan kepada Mahasiswa. Namun dengan banyaknya artikel yang ditawarkan setiap hari justru membuat mahasiswa kewalahan dan cenderung memilih artikel yang kurang sesuai dengan apa yang sebenarnya diinginkan. Untuk membantu menemukan informasi yang relevan, dikembangkanlah sistem rekomendasi. Sistem rekomendasi akan membantu mengestimasi nilai prediksi atau relevansi dari sebuah artikel dan kemudian membuat ranking atas artikel-artikel tersebut yang sesuai dengan minat pengguna. Teknik Content-based Filtering dipilih dalam pembuatan sistem rekomendasi pada penelitian ini. Dengan menggunakan dataset dari portal berita Kabar Informatika Universitas Amikom Yogyakarta, sistem rekomendasi Content-based Filtering pada penelitian ini mampu memberikan skor Recall@5 sekitar 73% dan Recall@10 sekitar 80%.

**Kata Kunci:** Artikel Berita; Content-based Filtering; Cosine Similarity; Sistem Rekomendasi; TF-IDF

**Abstract**—The population of active students in the Informatics Bachelor Program, Universitas Amikom Yogyakarta, in the odd semester of 2021 is 3,870. Efforts to track interest in the three concentration options were carried out early on through article literacy recommendations. Various articles are produced continuously and provided on an ongoing basis to students. However, the many articles offered daily make students overwhelmed and tend to choose articles that do not match what they want. To help solve this problem, recommender system is developed. A recommender system helps to estimate the prediction value or relevancy of an article and create a ranking according to user's taste. Content-based Filtering technique is used in this research. Using the dataset from Kabar Informatika news portal of University of Amikom Yogyakarta, the developed Content-based Filtering Recommendation System is able to produce Recall@5 score at around 73% and Recall@10 at around 80%.

**Keywords:** Content-based Filtering; Cosine Similarity; News Article; Recommender System; TF-IDF

## 1. PENDAHULUAN

Diambil dari Pangkalan Data Dikti 2022 [1], populasi mahasiswa aktif prodi Informatika pada semester Ganjil 2021 sebanyak 3870 orang. Upaya pelacakan minat terhadap tiga pilihan konsentrasi dilakukan sejak dini, melalui rekomendasi literasi artikel. Tiga konsentrasi tersebut meliputi Software Engineering & Big Data, Cloud Networking dan Multimedia yang kemudian dalam platform berita direpresentasikan sebagai kategori artikel. Ragam artikel diproduksi terus menerus dan diberikan secara berkelanjutan kepada Mahasiswa. Namun dengan banyaknya artikel yang ditawarkan setiap hari justru membuat mahasiswa kewalahan dan cenderung memilih artikel yang kurang sesuai dengan apa yang sebenarnya diinginkan.

Untuk membantu menemukan informasi yang relevan diantara banyaknya informasi yang ada, dikembangkanlah Sistem Rekomendasi [2]. Tujuan dari sistem rekomendasi adalah untuk memberikan konten (*item*) secara efektif dan berarti kepada pengguna yang aktif didalam *platform* [3]. Sistem Rekomendasi adalah seperangkat peralatan dan teknik yang memberikan saran kepada pengguna tentang *item* spesifik yang kemungkinan mereka minati. Sebuah sistem rekomendasi mencatat profil pengguna dan berdasarkan minat mereka, menyarankan sebuah produk atau layanan [4]. Saran yang diberikan bisa dalam *domain* apapun mulai dari saran layanan web apa yang digunakan hingga artikel berita untuk dibaca [5]–[7].

Terdapat dua tipe sistem rekomendasi, yang dipersonalisasi dan yang tidak dipersonalisasi. Sistem rekomendasi yang dipersonalisasi adalah sistem yang mana setiap pengguna mendapatkan saran yang berbeda, sedangkan sistem rekomendasi yang tidak dipersonalisasi semua pengguna mendapatkan saran yang sama [8]. Yang perlu dilakukan oleh sistem rekomendasi yang dipersonalisasi adalah dua hal, yaitu (i) mengestimasi nilai prediksi untuk sebuah item (ii) memberikan *ranking item-item* tersebut berdasarkan nilai prediksinya [9].

Terdapat beberapa pengelompokan dalam sistem rekomendasi. Beberapa yang populer antara lain *Content-based Filtering*, *Collaborative Filtering*, dan kombinasi antara keduanya yang disebut dengan *Hybrid* [10]. Untuk dapat memberikan manfaat kepada pengguna, sistem rekomendasi dituntut untuk bisa memprediksi item yang relevan dengan pengguna. Pada referensi berikut untuk mencapai tujuan tersebut diperlukan beberapapertimbangan. Diantaranya adalah pertimbangan dalam pemilihan teknik sistem rekomendasi yang disesuaikan dengan ketersediaan data dan konteks penerapannya [11]. Untuk dapat memberikan manfaat kepada pengguna, sistem rekomendasi dituntut untuk bisa memprediksi item yang relevan dengan pengguna. Pada referensi berikut [11] untuk mencapai tujuan tersebut diperlukan beberapapertimbangan. Diantaranya adalah pertimbangan dalam pemilihan teknik sistem rekomendasi yang disesuaikan dengan ketersediaan data dan konteks penerapannya.

Sistem rekomendasi artikel berita, pernah dikembangkan menggunakan *dice similarity*[12] dan *K-Nearest Neighbors(KNN)* [13]. Pendekatan lain dilakukan [14] melalui pembobotan *Implicit Social Trust* dan *Support Vector Regression* untuk memprediksi nilai topik sebuah artikel. Sedangkan mekanisme peringkasan berita telah dilakukan [15] menggunakan fitur pembobotan *TFIDF*. Penelitian ini berfokus pada penggunaan metode *Content-based Filtering* dalam pembuatan sistem rekomendasi untuk artikel berita. Metode ini akan mendeskripsikan item, dalam hal ini artikel, menggunakan kata kunci. Algoritma yang dipakai dalam metode ini memprediksi *item* berdasarkan apa yang disukai oleh pengguna diwaktu lampau. Dalam *Content-based Filtering*, sistem membandingkan profil dari pengguna dengan profil dari konten (item) lalu mencari item yang mempunyai kemiripan kemudian menyarakannya ke pengguna [10]. Algoritma *TF-IDF* dan *Cosine Similarity* dirancang sedemikian rupa dan dikemas dalam sebuah package python sehingga dapat kembali digunakan untuk keperluan berikutnya.

## 2. METODOLOGI PENELITIAN

### 2.1 Landasan Penelitian

#### a. Sistem Rekomendasi

Sistem Rekomendasi adalah seperangkat peralatan dan teknik yang dapat memprediksi nilai preferensi atau rating sebuah item dengan memanfaatkan informasi personal seseorang dan karakteristik item [16]. Sistem ini mampu memberikan saran kepada pengguna tentang *item* spesifik yang kemungkinan mereka minati. Sebuah sistem rekomendasi mencatat profil pengguna dan berdasarkan minat mereka, menyarankan sebuah produk atau layanan [4][17]. Saran yang diberikan bisa dalam *domain* apapun mulai dari saran layanan web apa yang digunakan hingga artikel berita untuk dibaca [5]–[7]. Tujuan dari sistem rekomendasi adalah untuk memberikan konten (*item*) yang efektif dan berarti kepada pengguna yang aktif didalam platform [3]. Sebuah sistem rekomendasi secara abstrak dianggap sebagai integrasi dari tiga komponen utama, yaitu, pengumpulan data, mesin rekomendasi, dan antarmuka pengguna [18].

#### b. Content-based Filtering

*Content-based Filtering* adalah sebuah pendekatan yang relatif umum dalam bidang *Information Retrieval*. Area riset seputar *Content-based Filtering* berfokus pada rekomendasi konten tekstual seperti laman web, buku, dan film [19][20]. *Content-based Filtering* memberikan rekomendasi *item* dengan membandingkan fitur dari *item* tersebut [9], [19].

#### c. Term Frequency – Inverse Document Frequency (TF-IDF)

*TF-IDF* adalah sebuah model statistik untuk mengevaluasi seberapa penting kata dalam kumpulan dokumen [21]. *TF-IDF* digunakan secara luas untuk pemilihan fitur dalam pemrosesan informasi teks. Algoritma ini terdiri dari dua bagian: (1) *Term Frequency* (TF) yang merepresentasikan frekuensi kemunculan sebuah kata dalam teks; (2) *Inverse Document Frequency* (IDF) mengukur seberapa penting sebuah kata yang diimbangi dengan seberapa sering kemunculan sebuah kata dalam keseluruhan data. Nilai signifikansi dari sebuah kata meningkat bersamaan dengan frekuensi kemunculan pada sebuah teks, namun berbanding terbalik dengan meningkatnya frekuensi kemunculan dalam kumpulan teks [22]. *TF-IDF* adalah algoritma skema pemberatan yang paling populer [23]. Berikut adalah formula dari *TF-IDF*:

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad (1)$$

$$idf(t, D) = \log \frac{N}{|\{d \in D: t \in d\}|} \quad (2)$$

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \quad (3)$$

#### d. Vector Space Model

*Vector Space Model* adalah metode untuk mempresentasikan konten dari item, dalam hal ini konten artikel, kedalam vektor *n*-dimensi dimana *n* adalah jumlah keseluruhan kata yang telah diekstrak dari konten. Setiap posisi dalam vektor memberikan nilai berat (relevansi) dari sebuah kata pada item atau pengguna dan dihitung menggunakan *TF-IDF* [9].

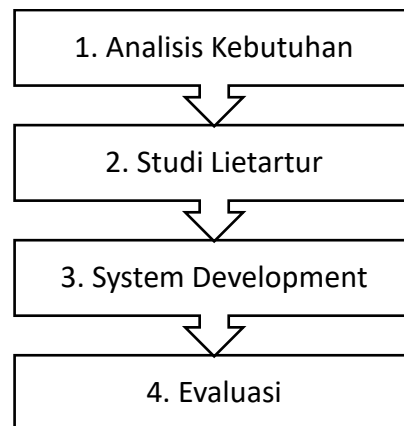
#### e. Cosine Similarity

Dua buah item dapat diukur nilai kemiripannya melalui beberapa cara, namun secara garis besar didefinisikan sebagai berikut: diberikan dua item,  $i_1$  dan  $i_2$ , tingkat kemiripan diantara keduanya ditulis dengan fungsi  $\text{sim}(i_1, i_2)$ [24]. Salah satu algoritmanya adalah *Cosine Similarity*. Secara matematik, setiap dokumen diproyeksikan sebagai sebuah vektor multi-dimensi, kemudian dihitung sudut cosine antara kedua vektor. Semakin kecil sudutnya maka semakin mirip dokumen tersebut. Pada sistem rekomendasi *Content-based Filtering*, *Cosine Similarity* diaplikasikan pada vektor fitur untuk mencari item memiliki kesamaan [9][25]. Berikut ini adalah formula dari *Cosine Similarity*:

$$sim(i_1, i_2) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (4)$$

## 2.2 Tahapan Penelitian

Tahapan penelitian secara garis besar ditunjukkan pada Gambar 1. Tahap pertama dimulai dari melakukan analisis kebutuhan dan ketersediaan data yang tersedia di portal berita Kabar Informatika dan melakukan studi literatur. Hal ini dilakukan dalam rangka untuk memilih metode sistem rekomendasi yang dapat memanfaatkan data yang tersedia dengan baik.

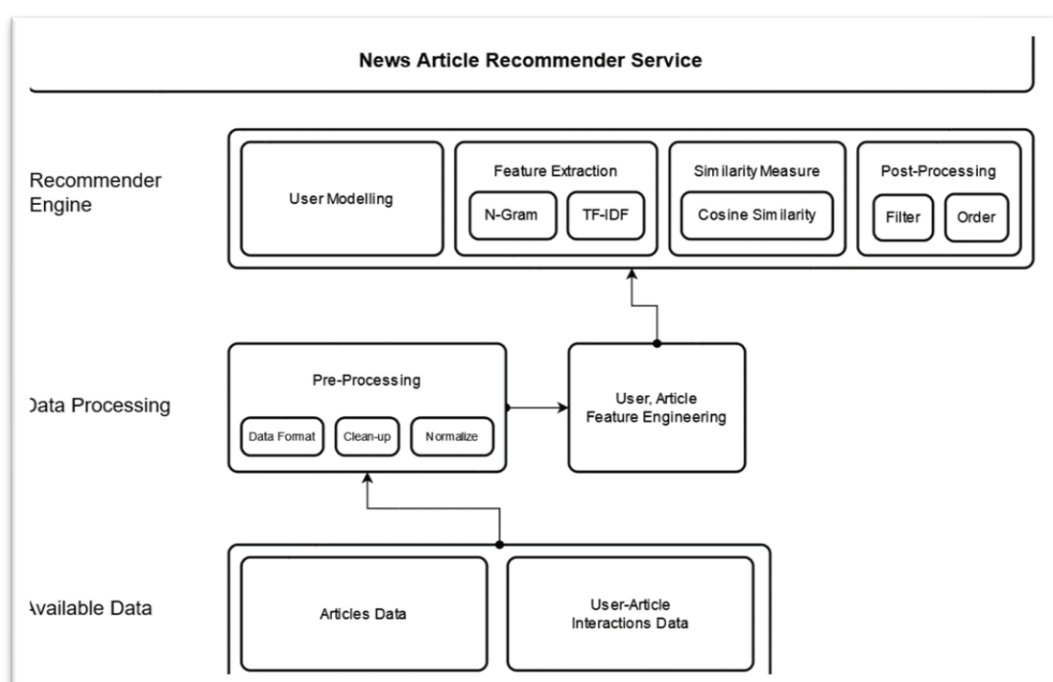


Gambar 1. Tahapan penelitian

Dataset yang digunakan berasal dari portal berita Kabar Informatika Universitas Amikom Yogyakarta. Dataset yang tersedia memiliki fitur cukup lengkap seperti judul, penulis, kategori, dan konten artikel itu sendiri yang sangat mendukung dalam proses pembuatan rekomendasi. Dataset ini cukup kecil dengan jumlah artikel sebanyak 22 dan jumlah data interaksi artikel sebanyak 345 yang berasal dari 20 pengguna.

Tahap kedua melakukan peninjauan literasi melalui pencarian buku serta penelitian serupa melalui portal Garuda dan Google Scholar. Dilanjutkan tahap ketiga *System Development* yang mencakup proses perancangan, pengembangan dan implementasi. Proses perancangan dilakukan dengan metode waterfall. Metode waterfall dipilih karena proses perancangan setiap tahapnya yang secara alami bergantung dengan tahap sebelumnya untuk mulai mengerjakan tahap selanjutnya.

Rancangan arsitektur package mesin rekomendasi yang dibangun ditunjukkan pada Gambar 2. Proses perancangan dimulai dari pengumpulan data, pre-processing data yang telah dikumpulkan, membuat mesin rekomendasi, dan diakhir dilakukan evaluasi. Tahap evaluasi akan dilakukan menggunakan metrik Recall.



Gambar 2. Arsitektur Sistem Rekomendasi

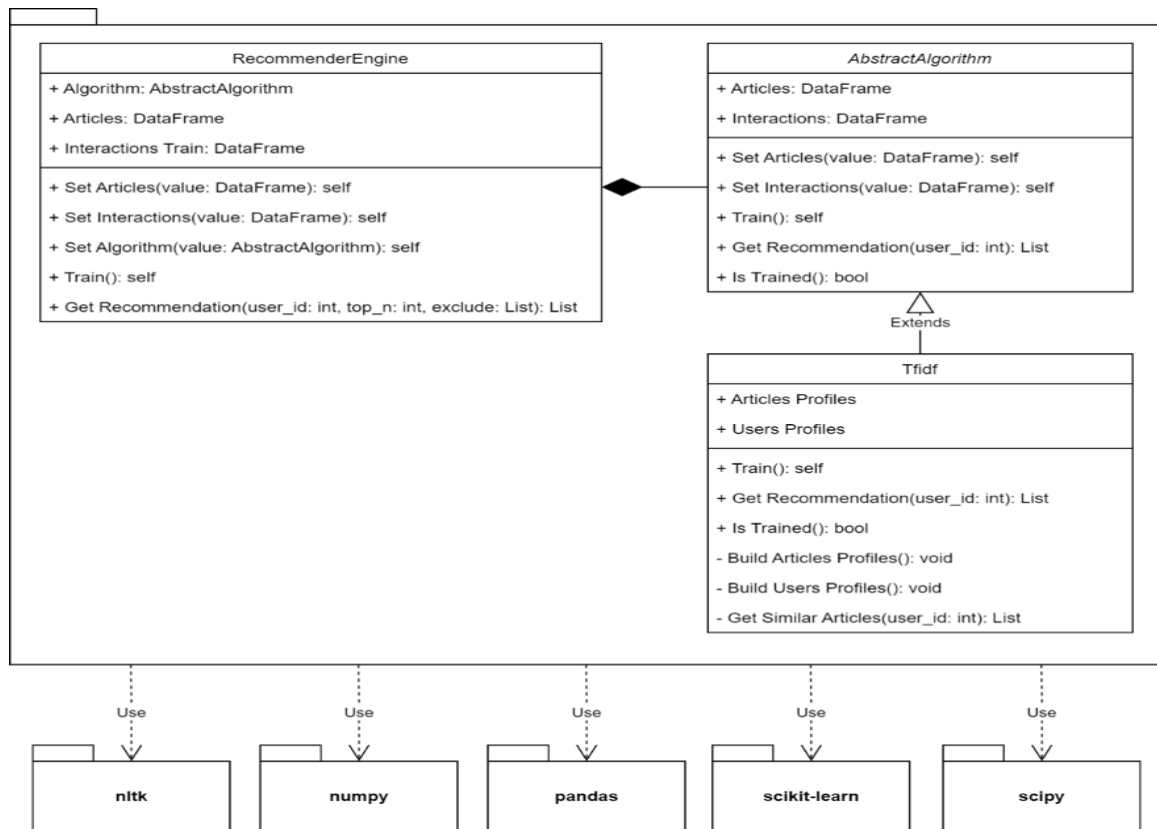
Semua elemen saling bekerja sama dalam setiap tahapan proses yang terjadi dimulai dari data mentah sampai dengan produksi rekomendasi artikel. Layer Available Data menyediakan data operasional dari platform berita ke layer di atasnya. Pada layer Data Processing, bagian Pre-Processing akan melakukan berbagai pemrosesan awal yang diperlukan agar data operasional dapat diproses. Selanjutnya bagian Feature Engineering mulai menyeleksi fitur-fitur yang tersedia pada data untuk kebutuhan pembuatan profil rekomendasi. Data yang telah diproses sampai tahap ini kemudian masuk ke layer Recommender Engine dimana data akan diproses menjadi rekomendasi. Secara garis besar terdapat tiga tahapan proses yang ditangani oleh empat bagian berbeda pada layer ini: (1) Membangun profil artikel dan pengguna; (2) Membuat rekomendasi; (3) Menyajikan data rekomendasi ke pengguna.

Komputer yang digunakan dalam penelitian berspesifikasi CPU Intel Core i7 9750H, RAM 8GB dan SSD 256GB. Pada proses penelitian menggunakan bahasa pemrograman Python dengan beberapa library yang terkait dalam topik, karena dalam penelitian menggunakan algoritma matematika menggunakan Library NumPy, dan beberapa library untuk menangani natural language Toolkit menggunakan NLTK. Dan dengan Python sebagai Enverimentnya, Untuk seluruh daftar perangkat lunak dan pustaka (library) yang digunakan ada pada tabel 1 di bawah:

**Tabel 1.** Daftar Perangkat Lunak

Nama	Versi
Windows 11	21H1
PyCharm Professional	2021.3.1
Anaconda	2021.11
Python	3.9
Beautiful Soup	4.10.0
NLTK	3.6.7
NumPy	1.22.0
Pandas	1.3.5
Requests	2.26.0
Scikit-Learn	1.0.2
SciPy	1.7.3
Unidecode	1.3.2

Pada penelitian ini, layer Recommender Engine diimplementasikan sebagai Package pada bahasa pemrograman Python. Package ini didesain dengan memanfaatkan prinsip-prinsip dalam paradigma Pemrograman Berorientasi Objek seperti Abstraksi, Enkapsulasi, Pewarisan, dan Polimorfisme seperti yang diilustrasikan pada Gambar 3 di bawah.



**Gambar 3.** Class Diagram Package Sistem Rekomendasi

*Package Recommender Engine* pada Gambar 3 diatas dapat digunakan secara terpisah tanpa bergantung pada layer-layer dibawahnya. Sumber data lain dapat menggunakan *Recommender Engine* ini dengan menyamakan struktur data yang digunakan. Dalam dataframe dibuat dalam 2 kolom untuk struktur data artikel dengan 2 variabel ID Artikel dengan data Int dan Fitur Artikel dengna type data string dapat di lihat pada tabel 2 di bawah.

**Tabel 2.** Struktur data artikel

DataFrame: Artikel		
Column	Type	Deskripsi
0	Int	ID Artikel
1	String	Fitur Artikel (Judul, Konten, dll)

Selanjutnya dibutuhkan tabel data Interaksi, yang di buat dalam 3 kolom di mana terdapat ID pengguna, ID artikel serta skor bobot nantinya, lihat tabel 3 di bawah.

**Tabel 3.** Struktur data interaksi

DataFrame: Interaksi Pengguna - Artikel			
Column	Type	Index	Deskripsi
user_id	Int String	True	ID Pengguna
article_id	Int	False	ID Artikel
weight	Float	False	Skor Bobot Interaksi Rata-rata

*Dataset* yang digunakan berasal dari portal berita Kabar Informatika Universitas Amikom Yogyakarta. *Dataset* yang tersedia memiliki fitur cukup lengkap seperti judul, penulis, kategori, dan konten artikel itu sendiri yang sangat mendukung dalam proses pembuatan rekomendasi. *Dataset* ini cukup kecil dengan jumlah artikel sebanyak 22 dan jumlah data interaksi artikel sebanyak 345 yang berasal dari 20 pengguna.

### 3. HASIL DAN PEMBAHASAN

Sistem rekomendasi yang dibuat adalah sistem rekomendasi top-N. Sistem rekomendasi ini akan membuat daftar rekomendasi artikel sebanyak N yang diurutkan berdasarkan seberapa relevan artikel tersebut dengan minat pengguna. Algoritma TF-IDF digunakan untuk menemukan kata-kata penting pada setiap artikel dan dilakukan penilaian pada setiap kata.

Proses perancangan sistem rekomendasi ini melalui beberapa tahap. Yang pertama, sumber data akan dikonsumsi melalui REST API yang memberikan data dalam format JSON. Kemudian data tersebut akan melalui tahap pre-processing. Pemrosesan yang ada di tahap ini berupa pembersihan data untuk menghilangkan semua kode formatting HTML yang ada. Setelah itu data akan melalui proses penghapusan stop word untuk meningkatkan efektivitas algoritma TF-IDF yang akan digunakan.

Kedua, data yang telah melalui tahap pre-processing akan mulai diproses dengan melibatkan algoritma TF-IDF (Term Frequency – Inverse Document Frequency). Algoritma TF-IDF digunakan untuk mencari kata-kata penting yang terdapat pada setiap artikel dengan memberikan skor untuk setiap kata. Contohnya jika kata “saya” frekuensi kemunculannya sangat tinggi di setiap artikel, maka algoritma ini akan memberikan skor rendah untuk kata tersebut di artikel yang memilikinya karena dianggap sebagai kata umum yang tidak memberikan informasi unik dari artikel. Sebaliknya jika kata “data mining” hanya sering muncul di beberapa artikel, maka algoritma ini akan memberikan skor tinggi. Hasil dari kalkulasi ini akan digunakan sebagai profil dari setiap artikel.

Ketiga adalah konstruksi profil minat pengguna dari data riwayat interaksi. Interaksi artikel yang terekam adalah interaksi membaca (Read), memberikan reaksi seperti menyukai artikel (React), dan membagikan artikel (Share). Masing-masing interaksi memiliki nilai atau bobot yang berbeda. Pada penelitian ini interaksi Read bernilai 1, React bernilai 2, dan Share bernilai 3. hasil uji data untuk riwayat data interaksi pengguna bisa di lihat pada tabel 4 di bawah

**Tabel 4.** Sampel data riwayat interaksi pengguna

user_id	article_id	event	weight
1	3	SHARE	3
1	3	COMMENT	2
1	4	SHARE	3
1	4	REACT	1
1	5	SHARE	3
1	5	COMMENT	2
1	6	COMMENT	2
1	6	REACT	1
1	11	REACT	1

1	13	COMMENT	2
1	16	SHARE	3
1	16	REACT	1
1	17	REACT	1
1	18	REACT	1
1	19	COMMENT	2
1	19	REACT	1
1	20	COMMENT	2
1	21	REACT	1

Data skor interaksi individu dari setiap artikel akan dijumlahkan dan dirata-rata. Pengguna dapat berinteraksi dengan artikel berkali-kali. Karena itu untuk menormalkan data, logaritma natural diaplikasikan ke jumlah total dari skor interaksi setiap artikelnya. Hasil dari tahap ini akan memberikan skor akhir ke tiap artikel yang menandakan seberapa tertarik pengguna dengan artikel tersebut. Hasil dari pengujian data hasil sampel data riwayat interaksi pengguna yang telah di rata rata berdasarkan dari hasil tabel 4 di atas dapat di lihat pada tabel 5 di bawah.

**Tabel 5.** Sampel data riwayat interaksi pengguna yang telah di rata-rata [2]

user_id	article_id	weight
1	3	4,584963
1	4	2,321928
1	5	2,584963
1	6	2,000000
1	11	1,000000
1	13	1,584963
1	16	2,321928
1	17	1,000000
1	18	1,000000
1	19	2,000000
1	20	1,584963
1	21	1,000000

Keempat, mencari kesamaan artikel dari profil yang telah dikonstruksi menggunakan Vector Space Model dengan pengukurannya menggunakan formula Cosine Similarity. Secara teknis Cosine Similarity mencari cosinus dari sudut antara dua vektor, semakin mirip arah sudutnya maka semakin mirip kedua vektor tersebut. Vektor dalam hal ini adalah profil artikel, lebih tepatnya data skor dari setiap kata dalam setiap artikel yang merupakan hasil dari kalkulasi TF-IDF. Dengan menggunakan Cosine Similarity, proses pembuatan rekomendasi dilakukan dengan mencari kemiripan artikel yang tersedia dengan artikel yang pernah berinteraksi dengan pengguna yang terekam dalam data riwayat interaksi.

Yang terakhir, daftar rekomendasi artikel yang dihasilkan pada tahap sebelumnya akan disesuaikan kembali sebelum disajikan kepada pengguna. Pemrosesan seperti cara pengurutan rekomendasi, seberapa banyak artikel yang disajikan, pengecualian artikel yang tidak ingin ditampilkan, dan semua pemrosesan lain yang diperlukan untuk mendukung penyajian rekomendasi akan dilakukan tahap ini.

Berikut ini adalah hasil pengujian sistem dalam pembuatan rekomendasi untuk beberapa pengguna yang berbeda yaitu pengguna 1 dan pengguna 2, hasil relevance dari pengguna 1 dan pengguna 2 dapat di lihat pada table 6 dan table 7 di bawah.

**Tabel 6.** Hasil rekomendasi untuk pengguna 1

id	title	relevance
5	Tips dan trik jago ngoding bagi pemula	0,510365
6	Mengenal Programming: Pengertian, Macam-Macam Bahasa Pemrograman, Jenis Profesi dan Manfaatnya	0,471760
3	Masih Gagal Paham Abstraction	0,405593
4	Perbedaan antara Front-end, Back-End dan Full-Stack Developer	0,396926
16	Artificial Intelligence in Business: The Importance & How AI Carries an Impact On Your Future...	0,360467
13	5 Kekurangan Programmer Pemula Jaman Now	0,332750
19	What is Netiquette? How important is Netiquette? What are the rules of Netiquette?	0,295908
20	Peluang Freelance Worker, sebagai Masa Depan Zillennial	0,254306
17	Big Data, Potensi Tambang Emas Masa Depan	0,224291
11	Zillennials Wajib Tahu, Inilah Startup di Indonesia yang Jadi Primadona Investasi Asia Tenggara	0,190918

**Tabel 7.** Hasil rekomendasi untuk pengguna 2

id	title	relevance
12	5 Alasan Mengapa Yogyakarta Menjadi Kota Developer-nya Startup	0,483614
13	5 Kekurangan Programmer Pemula Jaman Now	0,459938
16	Artificial Intelligence in Business: The Importance & How AI Carries an Impact On Your Future...	0,433289
11	Zillennials Wajib Tahu, Inilah Startup di Indonesia yang Jadi Primadona Investasi Asia Tenggara	0,382858
15	Mendulang Uang dengan Kecerdasan Buatan	0,358070
1	Membedah Ilmu Informatika Bersama Santri	0,355859
18	Artificial Intelligence Buatan Petruk, Menggemparkan Karangkadempel	0,336855
7	Swift Programming : Modular is Better	0,241970
5	Tips dan trik jago ngoding bagi pemula	0,208354
9	Simple MVP architecture for iOS App	0,194942

Proses evaluasi dilakukan menggunakan metrik pengukuran Recall. Cara kerja pengukuran metrik ini adalah dengan mengambil n-sampel artikel yang belum pernah berinteraksi dengan pengguna yang bersangkutan, kemudian meminta sistem rekomendasi yang sudah di-train dengan dataset training untuk membuat daftar rekomendasi top-n namun dengan menghilangkan artikel yang pernah berinteraksi dengan pengguna di dalam training set. Masing-masing artikel dalam data interaksi testing set kemudian digabungkan dengan sampel artikel yang sebelumnya didapatkan. Jika hasil rekomendasi juga merekomendasikan artikel yang berada dalam daftar gabungan artikel yang belum pernah berinteraksi tersebut, maka hasil rekomendasi dianggap Hit atau berhasil. Tabel 8 dibawah merupakan hasil dari evaluasi dengan matrik recall.

**Tabel 8.** Hasil evaluasi dengan metrik *Recall*

Pengguna	Total Artikel	Top 5 Hit	Top 10 Hit	Recall 5	Recall 10
1	2	2	2	1,00	1,00
2	2	2	2	1,00	1,00
3	2	1	1	0,50	0,50
4	2	2	2	1,00	1,00
5	3	1	2	0,33	0,67
6	2	1	1	0,50	0,50
7	3	2	3	0,67	1,00
8	3	2	2	0,67	0,67
9	3	3	3	1,00	1,00
10	2	1	1	0,50	0,50
11	2	2	2	1,00	1,00
12	2	1	1	0,50	0,50
13	3	2	2	0,67	0,67
14	2	1	1	0,50	0,50
15	2	2	2	1,00	1,00
16	2	2	2	1,00	1,00
17	2	1	1	0,50	0,50
18	2	2	2	1,00	1,00
19	2	1	2	0,50	1,00
20	2	2	2	1,00	1,00
<b>Global Recall</b>				<b>0,73</b>	<b>0,80</b>

Hasil evaluasi menunjukkan metrik Recall@5 memberikan skor 0,73 atau sekitar 73%, kemudian Recall@10 memberikan skor 0,80 atau sekitar 80%. Artinya persentase dari sistem rekomendasi ini memberikan rekomendasi artikel yang relevan dengan pengguna di dalam daftar top 5 sebesar 73% dan dalam daftar top 10 sebesar 80%.

#### 4. KESIMPULAN

Dengan dataset yang meskipun tidak memiliki banyak item namun memiliki fitur cukup lengkap masih sangat membantu sistem dalam membuat rekomendasi artikel yang sesuai dengan minat pengguna. Hasil evaluasi yang memberikan Recall@5 sekitar 73% dan Recall@10 sekitar 80% menunjukkan sistem dapat memberikan daftar rekomendasi dengan baik dan cukup relevan dengan apa yang pengguna minati. Meskipun demikian, keterbatasan data dapat diatasi dengan menambahkan, menggabungkan, atau menggantinya. Penggantian menjadi lebih mudah dikarenakan mesin rekomendasi yang tidak memiliki dependensi ke dataset yang digunakan. Tentu masih banyak aspek dalam penelitian ini yang dapat ditingkatkan untuk membantu memperbaiki performa dari sistem rekomendasi.

Selain dengan meningkatkan kuantitas dari artikel itu sendiri, terdapat banyak hal yang dapat ditingkatkan. Pada aspek tokenization artikel contohnya, token relevansi yang diberikan saat memberikan rekomendasi menunjukkan beberapa kata yang berasal dari potongan kode program yang digunakan dalam artikel.

## REFERENCES

- [1] Pangkalan Data Pendidikan Tinggi, “Jumlah Mahasiswa Prodi Informatika Universitas Amikom Yogyakarta Tahun 2019-2021 berdasarkan Pangkalan Data Pendidikan Tinggi,” 2022. [Online]. Available: [https://pddikti.kemdikbud.go.id/data\\_pt/QzJERjg3QzMtMUE0RC00RjFBLTIDREYtNERENEY1NzBDQUE1](https://pddikti.kemdikbud.go.id/data_pt/QzJERjg3QzMtMUE0RC00RjFBLTIDREYtNERENEY1NzBDQUE1).
- [2] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, “Recommender systems survey,” *Knowledge-Based Syst.*, vol. 46, pp. 109–132, Jul. 2013.
- [3] A. Gatzoura and M. Sanchez-Marre, “A Case-Based Recommendation Approach for Market Basket Data,” *IEEE Intell. Syst.*, vol. 30, no. 1, pp. 20–27, Jan. 2015.
- [4] Y. Wang, N. Stash, L. Aroyo, L. Hollink, and G. Schreiber, “Semantic Relations in Content-based Recommender Systems,” *Proc. fifth Int. Conf. Knowl. capture*, pp. 1–8, 2009.
- [5] Z. Cao, X. Qiao, S. Jiang, and X. Zhang, “An efficient knowledge-graph-based web service recommendation algorithm,” *Symmetry (Basel)*, vol. 11, no. 3, 2019.
- [6] X. Kong, M. Mao, W. Wang, J. Liu, and B. Xu, “VOPRec: Vector Representation Learning of Papers with Text Information and Structural Identity for Recommendation,” *IEEE Trans. Emerg. Top. Comput.*, vol. 9, no. 1, pp. 226–237, Jan. 2021.
- [7] V. Setty and K. Hose, “Event2Vec: Neural embeddings for news events,” *41st Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR 2018*, pp. 1013–1016, 2018.
- [8] K. Shah, A. Salunke, S. Dongare, and K. Antala, “Recommender systems: An overview of different approaches to recommendations,” in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017, pp. 1–4.
- [9] S. Kanwal, S. Nawaz, M. K. Malik, and Z. Nawaz, “A Review of Text-Based Recommendation Systems,” *IEEE Access*, vol. 9, pp. 31638–31661, 2021.
- [10] S. Gupta, “A Literature Review on Recommendation Systems,” *Int. Res. J. Eng. Technol.*, 2020.
- [11] F. Ricci, L. Rokach, and B. Shapira, *Recommender System handbook*. 2011.
- [12] S. Informatika and A. Polinema, “Implementasi Metode Dice Similarity Dalam Perancangan Sistem Rekomendasi Artikel Berita,” *Siap*, p. 2020, 2020.
- [13] A. S. Dharma, R. B. Basadena, A. Hutasoit, and R. R. Pangaribuan, “Sistem Rekomendasi Menggunakan Item-based Collaborative Filtering pada Konten Artikel Berita,” *Jurnalitio*, vol. 02, no. 01, 2021.
- [14] M. Widya Ningrum dan, “Implicit Social Trust Dan Support Vector Regression Untuk Sistem Rekomendasi Berita Implicit Social Trust and Support Vector Regression for News Recommender System,” vol. 3, no. 2, 2017.
- [15] N. K. Widyasanti, I. K. G. Darma Putra, and N. K. Dwi Rusjyanthi, “Seleksi Fitur Bobot Kata dengan Metode TFIDF untuk Ringkasan Bahasa Indonesia,” *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 6, no. 2, p. 119, 2018.
- [16] S. N. M. Pavan Kumar P., S. Vairachilai, Sirisha Potluri, *Recommender Systems Algorithms And Applications*. Oxon: CRC Press, 2021.
- [17] A. Masdalena et al., “Implementasi Metode Analytical Hierarchy Process Pada Sistem Pendukung Keputusan Rekomendasi Ikan Budidaya Berbasis Web,” vol. 4, no. 2, pp. 663–673, 2022.
- [18] S. Athalye, “Recommendation System for News Reader Recommendation System for News Reader A Project Presented to The Faculty of the Department of Computer Science San Jose State University In Partial Fulfillment Of the Requirements for the Degree Master of Science By S,” 2013.
- [19] M. D. Buhmann, *Encyclopedia of Machine Learning and Data Mining*. 2017.
- [20] E. T. Arifin, “Prediction Retweet Using User-Based and Content-Based with ANN-GA Classification Method,” vol. 4, no. 2, pp. 522–528, 2022.
- [21] M. Das, S. Kamalanathan, and P. Alphonse, “A Comparative Study on TF-IDF feature weighting method and its analysis using unstructured dataset,” *CEUR Workshop Proc.*, vol. 2870, pp. 98–107, 2021.
- [22] Z. Zhu, J. Liang, D. Li, H. Yu, and G. Liu, “Hot Topic Detection Based on a Refined TF-IDF Algorithm,” *IEEE Access*, vol. 7, pp. 26996–27007, 2019.
- [23] J. Beel, B. Gipp, S. Langer, and C. Breiteringer, “Research-paper recommender systems: a literature survey,” *Int. J. Digit. Libr.*, vol. 17, no. 4, pp. 305–338, Nov. 2016.
- [24] K. Falk, *Practical Reommender Systems*. Shelter Island, NY: Manning Publications Co., 2019.
- [25] I. G. Anugrah, “Penerapan Metode N-Gram dan Cosine Similarity Dalam Pencarian Pada Repositori Artikel Jurnal Publikasi,” *Build. Informatics, Technol. Sci.*, vol. 3, no. 3, pp. 275–284, 2021.