Journal of Informatics and Computer Engineering Research (JICER)
Vol. 2, No. 1, Juni 2025, hlm. 18-24
ISSN: 3089-6320

IMPLEMENTASI TF-IDF DAN KNN PADA REKOMENDASI JURNAL OTOMATIS

Jonathan Wijaya [1], Fernando Sugianto Putra [2], Hafiz Irsyad [3], Abdul Rahman [4]
[1,2,3] Informatika, Universitas Multi Data Palembang, Indonesia
[4] Elektro, Universitas Multi Data Palembang, Indonesia

Info Artikel

Riwayat artikel:
Received, (2 Juni 2025)
Revised, (17 Juni 2025)
Accepted, (23 Juni 2025)

Kata kunci:
Cosine distance;
Jurnal;
K-Nearest Neighbor;
Rekomendasi;
TF-IDF

ABSTRACT

In the rapidly evolving digital era, finding relevant journals has become a challenge for students, researchers, and academics. The easy access to journal publications has led to an increase in the number of available journals. Therefore, there is a need for a recommendation system based on related journals automatically using the TF-IDF and K-Nearest Neighbor (KNN) approaches with cosine distance. The goal is to enhance the efficiency and accuracy of journal searches. The first step is to process the text taken from the titles and abstracts of the journals into numerical vectors using TF-IDF to determine the importance of words in each document. Then, KNN is used to find the journals that are most similar to the specified journal based on the distance between TF-IDF vectors. Precision@3 is used to evaluate the results of the top three recommendations. The evaluation results show highly relevant recommendations, with a Precision@3 value of 1. This system has successfully improved the efficiency and accuracy of automatic journal searches for relevant content.

ABSTRAK

Dalam era digital yang terus berkembang, pencarian jurnal yang relevan menjadi tantangan bagi pelajar, peneliti dan akademisi. Mudahnya akses terhadap publikasi jurnal, membuat jumlah jurnal yang tersedia meningkat. Oleh karena itu, diperlukannya suatu implementasi rekomendasi berdasarkan jurnal terkait secara otomatis dengan pendekatan *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *K-Nearest-Neighbor* (KNN) dengan *cosine distance*. Tujuannya adalah untuk meningkatkan efisiensi dan akurasi pencarian jurnal. Tahap pertama adalah mengolah teks yang diambil dari judul dan abstrak jurnal menjadi vektor numerik menggunakan TF-IDF untuk mengetahui seberapa pentingnya kata di setiap dokumen. Setelah itu, KNN digunakan untuk mencari jurnal yang paling mirip dengan jurnal yang ditentukan berdasarkan jarak antar vektor TF-IDF. *Precision@3* digunakan untuk mengevaluasi hasil dari tiga rekomendasi teratas. Hasil evaluasi menunjukan rekomendasi yang sangat relevan, dengan nilai *Precision@3* mencapai
1. Implementasi ini berhasil meningkatkan efisiensi dan akurasi pencarian jurnal yang relevan secara otomatis.

Penulis Korespondensi:

Jonathan Wijaya
Informatika, Universitas Multi Data Palembang, Jl. Rajawali No.14, Palembang
Email: jonathanwijaya_2226250009@mhs.mdp.ac.id

1. PENDAHULUAN

Kata jurnal, sudah sangat familiar di telinga. Terutama bagi yang sedang menempuh pendidikan baik itu di sekolah maupun perguruan tinggi. Jurnal adalah sebuah publikasi atau catatan tertulis yang berisi catatan, penelitian, atau pemikiran seseorang atau sekelompok orang dalam bidang tertentu [1]. Jurnal sendiri bertujuan untuk menyebarluaskan pengetahuan dan hasil penelitian kepada masyarakat. Jurnal dapat dipublikasi dalam berbagai bentuk, seperti artikel, laporan penelitian, karya ilmiah dan masih banyak lagi [1]. Jurnal memiliki peran penting di dunia pendidikan. Dengan membaca jurnal, kita dapat meningkatkan dan memperluas pengetahuan kita. Tidak hanya itu, kita mungkin bisa menemukan hal baru dari hasil penelitian seseorang dalam jurnal tersebut. Jurnal juga dapat menjadi sumber referensi dan acuan bagi penulis baru yang ingin membuat sebuah penelitian atau sejenisnya. Dengan mencari dan membaca jurnal yang relevan dengan penelitian yang akan dilaksanakan, penulis bisa mendapat gambaran awal atau sekedar pengetahuan dasar dari penelitian yang akan dilakukan.

Dalam perkembangan sistem informasi yang sangat pesat ini telah membawa perubahan signifikan dalam dunia akademis, khususnya dalam publikasi ilmiah. Publikasi ilmiah sendiri sudah ada sejak peradaban Mesopotamia (8000 SM). Ditemukannya mesin cetak pada abad ke-16 mulai menarik ilmuwan dan pemikir untuk membagikan hasil pemikirannya ke seluruh dunia [2]. Di zaman sekarang publikasi dapat dilakukan melalui internet seperti SINTA, GARUDA, DOAJ, Google Scholar dan arxiv. Hal ini tentu mempermudah para penulis untuk mempublikasi jurnal mereka.

Namun, seiring dengan mudahnya untuk mempublikasi jurnal dan semakin luasnya akses terhadap internet, jumlah jurnal yang tersedia juga meningkat secara signifikan. Ribuan bahkan jutaan jurnal ilmiah diterbitkan setiap tahunnya di berbagai bidang. Banyaknya jumlah jurnal ini tentu membawa tantangan tersendiri, terutama dalam memberikan rekomendasi jurnal yang relevan dan sesuai dengan kebutuhan pengguna. Para peneliti, mahasiswa, dan akademisi seringkali membutuhkan sistem yang dapat secara otomatis merekomendasikan jurnal lain yang berkaitan dengan jurnal yang sedang mereka baca. Tanpa sistem rekomendasi yang tepat, mereka harus mencari secara manual jurnal-jurnal terkait, yang dapat memakan waktu lama dan tidak efisien, terlebih jika mereka tidak memiliki pemahaman konteks atau kata kunci yang sesuai.

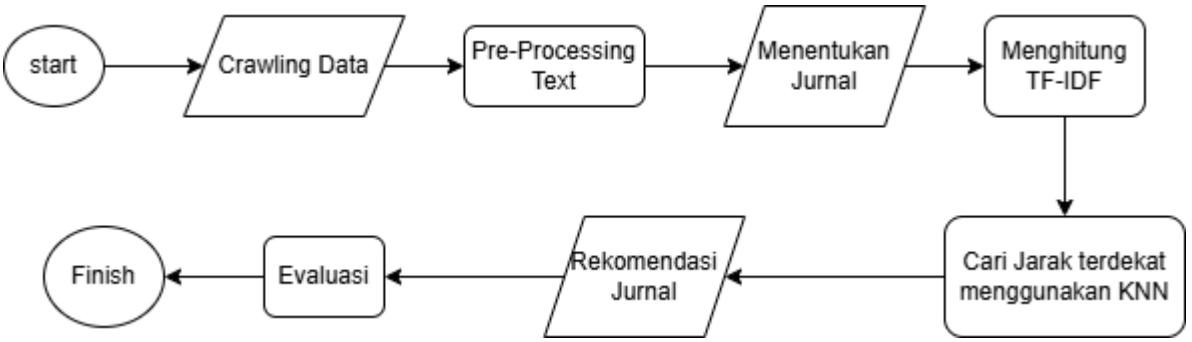
Pada penelitian terdahulu, dikembangkan sebuah model untuk meningkatkan akurasi sistem rekomendasi pada platform streaming . Algoritma yang digunakan adalah *collaborative filtering* dengan mengembangkan algoritma *K-Nearest Neighbor* (KNN) yang adaptif. Kekurangan dari *collaborative filtering* adalah *cold start*. Sistem kesulitan memberikan rekomendasi untuk item baru atau pengguna tanpa riwayat interaksi sebelumnya. Untuk mengatasi masalah tersebut, diterapkanlah algoritma KNN berdasarkan metode MAE, RMSE, MAP dan NDCG. Hasil pengujiannya memperoleh nilai MAE sebesar 0.810, RMSE 1.037, MAP
0.129 dan NDCG 0.167 [3]. Nilai ini lebih baik jika dibandingkan dengan collaborative filtering. Dengan demikian, algoritma KNN adaptif tidak hanya mengatasi kekurangan collaborative filtering, namun juga secara signifikan meningkatkan akurasi dan relevansi dari rekomendasi yang dihasilkan.

Berdasarkan latar belakang dan penelitian terdahulu, untuk mengatasi permasalahan tersebut, dibutuhkan sebuah implementasi yang dapat membantu pengguna dalam menemukan jurnal yang relevan secara otomatis. Salah satu solusi yang dapat diterapkan adalah dengan membangun rekomendasi jurnal menggunakan pendekatan *Term Frequency - Inverse Document Frequency* TF-IDF dan *K-Nearest Neighbor* (KNN). TF-IDF berfungsi untuk mengevaluasi seberapa penting dan membersihkan kata dari suatu dokumen untuk memudahkan dalam mengelola data untuk tahap selanjutnya. KNN adalah suatu metode yang menggunakan algoritma *supervised*, dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada KNN yang bertujuan untuk mengklasifikasi objek baru berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Memungkinkan representasi dokumen (dalam hal ini jurnal) dalam bentuk vektor berdasarkan kata-kata yang dikandungnya, sehingga kemiripan antara jurnal dapat dihitung secara matematis. Pendekatan ini tidak hanya mempercepat proses pencarian jurnal, tetapi juga meningkatkan akurasi dan relevansi hasil yang diperoleh.

2. METODE

Pendekatan yang digunakan dalam penelitian ini adalah kuantitatif dengan teknik pengumpulan data studi literatur. Metode penelitian kuantitatif merupakan pengumpulan dan analisis data numerik dengan variabel kontrol yang memungkinkan peneliti untuk menyelidiki fenomena dan hubungan antara variabel dengan pendekatan ilmiah yang terstruktur [4]. Studi literatur yaitu pengkajian data dari berbagai buku referensi serta hasil penelitian sebelumnya yang relevan dengan penelitian untuk mendapatkan landasan teori dari masalah yang akan diteliti [5] Tujuan dari metode penelitian ini adalah membangun sebuah implementasi rekomendasi jurnal berdasarkan judul yang sudah ditentukan oleh pembaca dengan pemrosesan teks TF-IDF dan KNN untuk mencari rekomendasi yang optimal.

Langkah - langkah dari metode ini adalah crawling dataset, pemrosesan teks dengan TF-IDF, mencari jarak terdekat dengan KNN, rekomendasi, dan evaluasi.



Gambar 1.Tahap Metode Penelitian

Proses Crawling Data diambil dari situs arxiv [6] yang diambil menggunakan API. {max_result} akan diisi dengan berapa banyak jumlah data yang akan diambil. Jumlah data yang diambil dapat ditentukan sendiri, karena jumlah jurnal yang sudah diterbitkan pada situs tersebut terlampau banyak. Disini, peneliti menetapkan 200 data jurnal kategori computer science - Artificial Intelligence yang akan diambil dan dilakukan tahap pra-pemrosesan. Data yang diambil dari jurnal tersebut adalah judul dan abstrak. Menurut peneliti, kedua data tersebut sangat relevan untuk mendukung sebuah sistem rekomendasi.

Pre-processing text bertujuan untuk menghilangkan karakter-karakter tertentu seperti tanda baca, simbol dan tokenize untuk mengubah kalimat/dokumen menjadi kata. Mengubah semua huruf pada teks menjadi huruf kecil dan menghapus kata stopword dengan bahasa inggris. Tujuan menghilangkan *stopword* ini yaitu menghilangkan teks yang tidak berhubungan dengan analisa sentimen [7].

Dalam tahap menentukan jurnal, user memilih jurnal diinginkan. Dari jurnal yang sudah dipilih, jurnal tersebut akan menjadi acuan untuk rekomendasi jurnal lainnya berdasarkan judul dan abstrak. Setelah didapatkan judul dan abstrak, kedua data tersebut akan diambil dan dilanjutkan ke tahap selanjutnya.

Tahap pemrosesan TF-IDF digunakan untuk mengkonversi teks ke dalam sebuah vektor numerik. metode TF-IDF memperhitungkan dua faktor penting, Term Frequency (TF) untuk mengukur seberapa sering suatu kata muncul dalam sebuah dokumen. *Inverse Document Frequency (IDF)* untuk mengukur seberapa penting suatu kata dalam konteks koleksi dokumen yang lebih besar. Untuk mencari nilai dari TF, IDF dan TF-IDF menggunakan persamaan (1)(2)(3)..

$$\text{TF} = \frac{Number\ of\ times\ a\ word\ "X"appears\ in\ a\ Document}{Number\ of\ words\ present\ in\ a\ Document} \quad (1)$$

$$\text{IDF} = \log\left(\frac{Number\ of\ Documents\ present\ in\ a\ Corpus}{Number\ of\ Documents\ where\ word\ "X"\ has\ appeared}\right) \quad (2)$$

$$\text{TF IDF} = \text{TF} * \text{IDF} \quad (3)$$

Setelah melalui tahap TF-IDF, data yang sudah bersih dapat digunakan untuk masuk ke dalam proses KNN. Metode KNN digunakan untuk mencari jarak terdekat dari jurnal yang sudah dipilih.. Untuk mencari nilai *Cosine Distance* pada KNN digunakan persamaan (4).

$$\textbf{Cosine Distance(A,B) = 1 -} \frac{A-B}{||A||\,||B||} \quad (4)$$

Setelah ditemukan jarak terdekat dari proses KNN, model akan memberikan rekomendasi jurnal lainnya yang mirip dengan jurnal yang sudah dipilih di awal oleh user. Model KNN akan memberikan 10 rekomendasi jurnal lainnya yang mirip berdasarkan judul dan abstrak.

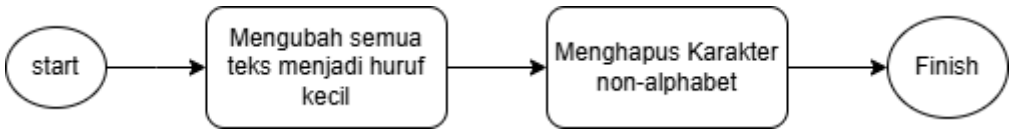
Evaluasi kinerja model rekomendasi menggunakan *Precision@K* yang menghitung akurasi implementasi [7]. Pengujian dilakukan terhadap query yang relevan terhadap hasil dari 10 rekomendasi jurnal dengan menggunakan KNN dengan K=3. 3 hasil teratas berdasarkan jarak yang paling dekat akan dihitung dengan *Precision@K* untuk melihat relevansi dari ke 3 hasil rekomendasi dengan jurnal yang dipilih. Rumus dari *Precision@K* adalah sebagai berikut:

$$Precision\ @K = \frac{r}{k}$$

Keterangan:
r = Jumlah dokumen relevan pada K dokumen teratas
K = Threshold peringkat (5)

3. HASIL DAN PEMBAHASAN

Tahap pertama yang dilakukan adalah mengcrawling data dari website arxiv untuk akan diambil judul dan abstraknya. Judul dan abstrak kemudian digabungkan menjadi sebuah kesatuan teks. Setelah itu, teks akan di filter yang berfungsi untuk mengubah semua huruf menjadi huruf kecil dan menghapus karakter non-alfabet, seperti angka, simbol, dan tanda baca. Tujuan dari pembersihan teks ini tentu untuk meningkatkan kinerja model dan fokus pada kata kunci yang relevan. Proses pembersihan text disajikan pada Gambar 2.



Gambar 2. Tahap Pembersihan Teks

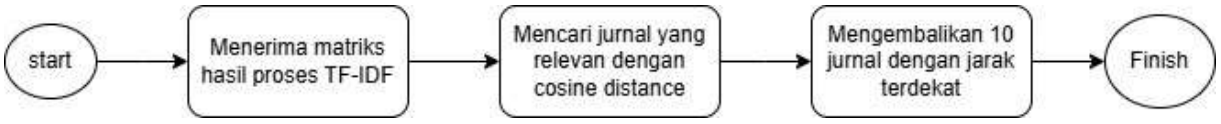
Selanjutnya, dilanjutkan ke dalam tahap menghitung TF-IDF, dengan mengubah teks yang telah dibersihkan menjadi representasi numerik. Teks akan diubah menjadi matriks TF-IDF yang setiap barisnya mewakili jurnal dan tiap kolom mewakili kata yang sudah dibersihkan dalam suatu jurnal. Fungsi dari *Term Frequency* (TF) adalah mengukur seberapa sering suatu kata muncul dalam suatu dokumen. *Inverse Document Frequency* (IDF) berfungsi untuk mengukur seberapa penting suatu kata dalam koleksi dokumen yang lebih besar. Dalam metode TF-IDF akan menghasilkan bobot kata untuk tiap kata di dalam sebuah dokumen yang menggambarkan tingkat pentingnya suatu kata dalam dokumen tersebut untuk dibandingkan dengan koleksi dokumen yang lebih besar [7].

Tabel 1. Hasil TF-IDF untuk 3 jurnal pertama

	ability	able	accepts	according	accounts	…	walras	wide	work	write	written
0	0	0	0	0	0	…	0	0	0	0	0
1	0	0	0	0.1009	0	…	0.1009	0	0	0	0
2	0	0	0	0	0.0836	…	0	0	0.0836	0	0

Hasil dari TF-IDF menunjukan tingkat pentingnya suatu kata di dalam satu dokumen, terdapat 2 hasil yang dapat dilihat pada tabel yaitu 0.1009 dan 0.0836. Semakin tinggi nilai yang dihasilkan, maka semakin penting kata tersebut di dalam suatu dokumen. Dalam tabel ini, nilai 0.1009 lebih tinggi daripada 0.0836. Dengan ini, kata ‘according’ dan ‘walras’ adalah kata yang sangat penting dalam dokumen 1.

KNN digunakan untuk mencari jurnal yang paling mirip dengan membandingkan vektor - vektor yang sudah dihasilkan pada proses TF-IDF sebelumnya. Pada penelitian ini, *Cosine Distance* digunakan untuk mencari jurnal yang paling mirip berdasarkan jarak *(distances)* dari jurnal yang ditentukan. 10 Jurnal akan dibandingkan dan mengambil 3 jurnal teratas dengan jarak yang paling dekat untuk ditampilkan ke pengguna dan menjadi rekomendasi. Proses KNN dengan *cosine distance* disajikan pada gambar 3

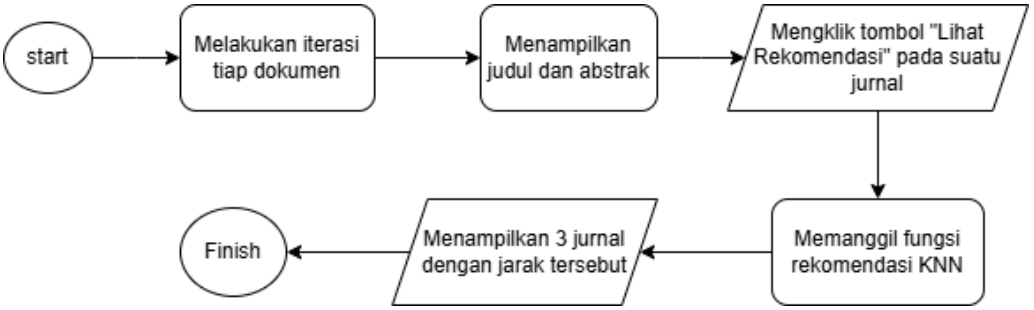


Gambar 3. Proses KNN dengan *Cosine Distance*

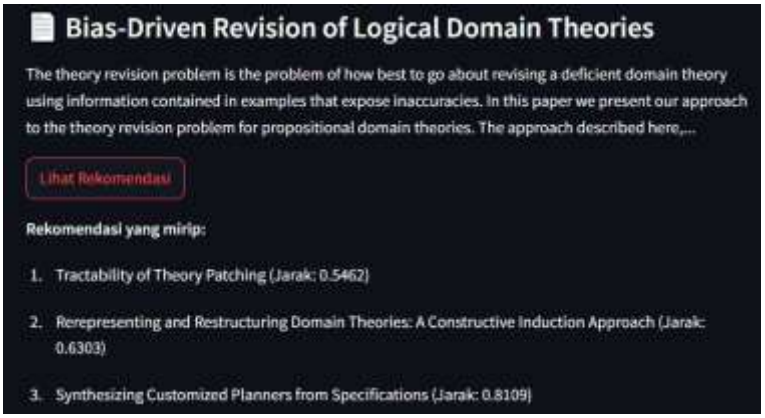
fungsi KNN akan menerima matriks hasil proses TF-IDF dan menggunakannya untuk mencari jurnal yang memiliki kemiripan paling tinggi dari jurnal yang ditentukan. Fungsi ini akan mengembalikan dua nilai yaitu indeks jurnal yang mirip dan jarak *(distances)* dengan rentang 0-1 dimana semakin kecil rentangnya, semakin relevan jurnal tersebut. knn.fit(tfidf_matrix) berfungsi untuk melatih model knn dengan data yang diberikan. Setelah dilatih, fungsi ini dapat mencari jurnal yang relevan dari jurnal yang diinput dengan metode *cosine distance.*

Hasil data crawling ditampilkan dalam sebuah UI Streamlit. Ada 200 jurnal yang ditampilkan dengan looping yang berisi judul dan abstraknya. Untuk melihat hasil rekomendasi, pengguna dapat mengklik tombol "Lihat Rekomendasi" pada akhir bagian dari tiap jurnal (di bawah abstrak). Ketika tombol ini diklik, model akan menjalankan fungsi KNN untuk mencari 10 jurnal dengan jarak terdekat. Model akan menampilkan 3 jurnal yang memiliki kemiripan dan jarak paling dekat dengan jurnal yang dipilih berdasarkan proses TF-IDF, KNN dan perhitungan jarak dengan cosine distance. Proses menampilkan hasil rekomendasi dapat dilihat pada gambar 4. Dalam contoh ini, jurnal dengan Bias-Driven Revision of Logical Domain Theoris digunakan sebagai jurnal acuan. Setelah diproses, implementasi ini akan menampilkan 3 jurnal relevan teratas, yaitu:

1. Tracibility of Theroy Pathcing (dengan jarak 0.54)
2. Rerepresenting and Restructuring Domain Theories: A Constructive Induction Approach (dengan jarak 0.63)
3. Syenthesizing Customized Planners from Spesifications (dengan jarak 0.81)

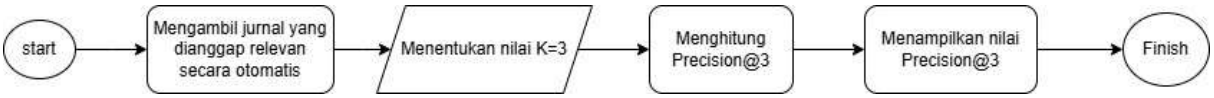


Gambar 4. Proses Penampilan Rekomendasi Jurnal

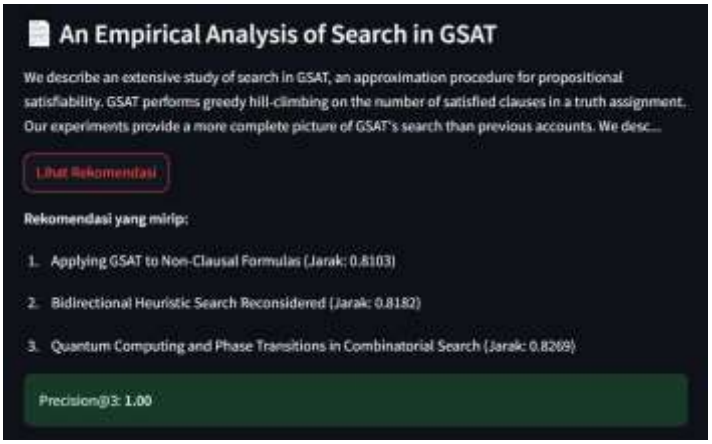


Gambar 5. Hasil Rekomendasi dengan UI *Streamlit*

Evaluasi *Precision@K* digunakan untuk mengukur seberapa relevan jurnal rekomendasi dari hasil KNN dengan *cosine distance*. Dalam prosesnya, nilai K ditentukan, yaitu 3, yang artinya 3 jurnal rekomendasi teratas yang didapat melalui metode KNN dan *Cosine Distance* akan dibandingkan dengan 3 jurnal yang relevan yang didapat melalui fungsi get_relevant_indices_auto. Fungsi ini akan mengembalikan nilai dengan rentang 0-1, dimana semakin tinggi nilainya, semakin relevan hasil rekomendasinya. Tujuan dari proses ini adalah untuk mengukur seberapa baik kualitas rekomendasi yang dihasilkan dari metode KNN dan *Cosine Distance.* Semakin tinggi nilai Precision, maka semakin relevan rekomendasi yang diberikan. Pada jurnal An Empirical Analysis of Search of GSAT menghasilkan nilai *Precision@3* sebesar 1. Hasil ini menunjukan bahwa, 3 hasil rekomendasi jurnal teratas memiliki relevansi yang sangat tinggi dengan jurnal terkait. Hasil ini dapat dilihat pada gambar 7.



Gambar 6. Proses perhitungan *Precision@K*



Gambar 7. Hasil *Precision@3*

4. KESIMPULAN

Berdasarkan hasil implementasi rekomendasi jurnal dengan pendekatan TF-IDF dan KNN dengan *Cosine Distance* dapat disimpulkan bahawa:

1. Proses pembersihan text dengan mengubah menjadi huruf kecil (*lowercase*), penghapusan karakter non-alfabet dan penghapusan kata yang tidak diperlukan (*stopword*) dapat berjalan dengan sangat baik dan mampu meningkatkan kinerja model untuk tahapan berikutnya
2. Pendekatan TF-IDF dan *K-Nearest Neighbor* (KNN) dengan *Cosine Distance* sudah sangat baik dalam memberikan rekomendasi jurnal. Pendekatan ini mampu memberikan rekomendasi jurnal yang relevan berdasarkan judul dan abstrak dengan menghitung jarak kemiripan antar jurnal.
3. Hal ini didukung dengan evaluasi menggunakan metode *Precision@K* dengan nilai K adalah 3. Hasil dari Precision ini adalah 1, yang menunjukan bahwa 3 rekomendasi jurnal teratas dengan jarak terdekat yang dihasilkan melalui proses TF-IDF dan KNN dengan *Cosine Distance* memiliki relevansi yang sangat tinggi dengan jurnal terkait.

REFERENSI

[1] "Apa Itu Jurnal? Pengertian, Jenis, dan Fungsi Jurnal," *e-Ujian*, [Online]. Available: https://e-ujian.id/apa-itu-jurnal-pengertian-jenis-dan-fungsi-jurnal.

[2] C. Mintardjo, "Publikasi pada Jurnal Ilmiah: Dari Era de Scavans Sampai Open Access Digital," *Kompasiana*, [Online]. Available: https://www.kompasiana.com/christoffelmintardjo9314/5f4db334d541df6fc4797eb2/publikasi-pada-jurnal-ilmiah-dari-era-de-scavans-sampai-open-access-digital.

[3] L. V. Nguyen, Q.-T. Vo, and T.-H. Nguyen, "Adaptive KNN-Based Extended Collaborative Filtering Recommendation Services," *Big Data Cogn. Comput.*, vol. 7, no. 2, pp. 106, 2023. [Online]. Available: https://www.mdpi.com/2504-2289/7/2/106/pdf?version=1685519525.

[4] R. A. Siroj, W. Afgani, F. Fatimah, D. Septaria, and G. Zahira, "Metode Penelitian Kuantitatif Pendekatan Ilmiah untuk Analisis Data," *Jurnal Review Pendidikan dan Pengajaran*, vol. 7, no. 3, pp. 11279-11289, 2024. [Online]. Available: https://journal.universitaspahlawan.ac.id/index.php/jrpp/article/view/32467/21663.

[5] Munib, A., & Wulandari, F. (2021). *Studi literatur: Efektivitas model kooperatif tipe Course Review Horay dalam pembelajaran IPA di sekolah dasar*. Jurnal Pendidikan Dasar Nusantara, 7(1), 160–170/ [Online]. Available:

https://ojs.unpkediri.ac.id/index.php/pgsd/article/download/16154/2176/

[6] "Search results from arXiv: AI," *arXiv*, [Online]. Available: http://export.arxiv.org/api/query?search_query=cat:cs.AI&start=0&max_results={max_results}.

[7] A. T. Jaka H, "Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining," *Jurnal Teknik Informatika*, vol. 2, no. 3, pp. 130-140, 2021. [Online]. Available: https://journal.upgris.ac.id/index.php/JIU/article/view/804.

[8] D. Septiani and I. Isabela, "Analisis Term Frequency Inverse Document Frequency (TF-IDF) dalam Temu Kembali Informasi pada Dokumen Teks," *SINTESIA: Jurnal Sistem dan Teknologi Informasi Indonesia*, vol. 1, no. 2, pp. 81-88, Mar. 2022, e-ISSN 2807-9108. [Online]. Available: https://journal.unj.ac.id/unj/index.php/SINTESIA/article/view/39364/15869.

[9] I. M. Ayudita, I. Indriati, and P. P. Adikara, "Sistem Pencarian Jurnal Ilmiah Cross Language dengan Metode Vector Space Model (VSM)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 12, pp. 6837-6841, Dec. 2018. [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/3765/1487/26454.