JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS)
E-ISSN : 2686-3359

Hal. 1807-1816
Vol. 7; No. 4
November 2025

PERBANDINGAN KINERJA *TF-IDF* DAN COUNT VECTORIZATION PADA SISTEM REKOMENDASI JUDUL SKRIPSI BERBASIS *CONTENT-BASED FILTERING*

Muhammad Arrafu Mazta[1*], Edi Saputra[2], Muhammad Razi A[3]
[1,2,3]Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Jambi
email: rafumazta@gmail.com[1*]

Abstrak: Penelitian ini bertujuan membandingkan dua skema representasi teks, *TF-IDF* dan *Count Vectorizer*, untuk membangun sistem rekomendasi judul skripsi berbasis *content-based filtering* pada repository Universitas Jambi. Kedua metode dipilih karena mewakili dua pendekatan pembobotan yang berbeda, *TF-IDF* menonjolkan istilah yang penting pada korpus sehingga cocok membedakan topik, sedangkan *Count Vectorizer* hanya berdasarkan frekuensi kemunculan kata dalam suatu dokumen tanpa mempertimbangkan sebarannya di korpus. Data berupa judul dan abstrak diperoleh melalui web scraping, kemudian diproses dengan deteksi bahasa, penghapusan *stop-word*, *stemming*, dan pembersihan teks. Untuk mengatasi ketiadaan label, dilakukan klasterisasi menggunakan *HDBSCAN* guna menghasilkan label tematik sementara, lalu subset berlabel (347 dokumen) dibagi menjadi 80% data latih dan 20% data uji dan dievaluasi menggunakan *K-Nearest Neighbors* dengan metrik *accuracy*, *precision*, *recall*, *F1-score*, serta analisis *confusion matrix*. Hasil menunjukkan kombinasi *TF-IDF* + *K-Nearest Neighbors* (k = 7) mencapai akurasi 98,57%, presisi 99,05%, *recall* 98,57%, dan *F1-score* 98,48%, melampaui *Count Vectorizer* yang tertinggi pada akurasi 94,29%. Prototipe *Streamlit* sebagai *proof of concept* menunjukkan bahwa *TF-IDF* menghasilkan rekomendasi yang lebih relevan dan efisien untuk penemuan skripsi di repository Universitas Jambi.

Kata Kunci : *Content-based filtering*, *K-Nearest Neighbor*, *Repository* skripsi.

PENDAHULUAN

Skripsi merupakan salah satu jenis karya ilmiah yang memiliki nilai strategis sebagai sumber referensi akademik, terutama karena memuat uraian teknis yang mendalam seperti metodologi, proses implementasi, serta detail analisis data yang sering kali tidak dijumpai dalam artikel jurnal. Di Universitas Jambi, *repository* daring telah menyimpan ribuan judul skripsi dari berbagai jurusan maupun program studi yang dilengkapi dengan abstraknya. Sayangnya, sistem pencarian dalam *repository* tersebut hingga saat ini masih mengandalkan metode pencocokan kata kunci secara konvensional yang belum mampu mengakomodasi pencarian berbasis makna atau relevansi semantik. Hal ini menyebabkan pencarian referensi menjadi tidak efisien dan menyulitkan mahasiswa atau peneliti dalam menemukan skripsi yang sesuai dengan topik atau minat penelitian mereka.

Untuk mengatasi permasalahan tersebut, pendekatan *content-based filtering* dapat diimplementasikan dalam pengembangan sistem rekomendasi judul skripsi. Pendekatan ini bekerja dengan mengukur kemiripan antar dokumen berdasarkan isi atau kontennya. Dua metode representasi teks yang banyak digunakan dalam pendekatan ini adalah *Term Frequency–Inverse Document Frequency* (*TF-IDF*) dan *Count Vectorization*. *TF-IDF* dikenal efektif dalam menyoroti kata-kata kunci penting dengan memberikan bobot lebih besar terhadap kata-kata yang jarang muncul dalam korpus, sehingga cocok digunakan dalam sistem rekomendasi berbasis teks [1]. Sementara itu, *Count Vectorization* menawarkan pendekatan yang lebih sederhana dengan hanya menghitung frekuensi kata, namun masih relevan terutama dalam dokumen pendek seperti judul dan abstrak.

Perbandingan performa kedua algoritma ini penting dilakukan guna mengetahui efektivitas masing-masing dalam konteks rekomendasi judul skripsi. Dalam penelitian ini, dilakukan proses evaluasi menggunakan algoritma *K-Nearest Neighbor* (*KNN*) sebagai metode klasifikasi dan *Confusion Matrix* untuk mengukur metrik evaluasi seperti akurasi, presisi, *recall*, dan *F1-score*. Untuk mengatasi tantangan pelabelan data yang besar dan tidak terstruktur, digunakan teknik *clustering* menggunakan *HDBSCAN* untuk mengelompokkan data berdasarkan kesamaan konten, yang kemudian digunakan sebagai dasar klasifikasi.

Terdapat beberapa penelitian terdahulu yang mengadopsi pendekatan serupa pada konteks berbeda. Penelitian oleh Zaynurroyhan *et al.* (2023) membandingkan *TF-IDF* dan *Count Vectorization* dalam sistem rekomendasi berbasis konten dan menunjukkan bahwa *TF-IDF* memiliki akurasi yang lebih tinggi dalam memberikan rekomendasi [1]. Hersianty *et al.* (2025) memanfaatkan *TF-IDF* untuk membangun sistem rekomendasi lowongan kerja dan menemukan bahwa metode ini efektif dalam menyarankan hasil sesuai preferensi pengguna [2]. Penelitian lain oleh Pradana *et al.* (2022) menerapkan *TF-IDF* dan *Cosine Similarity* untuk merekomendasikan kegiatan ekstrakurikuler kepada siswa, dengan hasil evaluasi menunjukkan relevansi yang tinggi [3]. Sementara itu, Ridwansyah *et al.* (2024) menggunakan *TF-IDF* dalam sistem rekomendasi produk digital dan berhasil menunjukkan performa yang baik, meskipun tidak membandingkannya langsung dengan algoritma lain [4].

Meskipun berbagai penelitian telah menunjukkan efektivitas penggunaan *TF-IDF* dalam sistem rekomendasi, kajian yang secara langsung membandingkannya dengan *Count Vectorization* dalam konteks skripsi, khususnya berbasis data dari *repository* akademik seperti Universitas Jambi, masih jarang dilakukan. Oleh karena itu, penelitian ini bertujuan

untuk mengembangkan dan mengevaluasi sistem rekomendasi judul skripsi berbasis *content-based filtering* dengan membandingkan kedua algoritma tersebut guna meningkatkan efisiensi dan ketepatan pencarian referensi akademik.

TINJAUAN PUSTAKA

Content-Based Filtering (*CBF*) merupakan metode dalam sistem rekomendasi yang menyarankan *item* kepada pengguna berdasarkan kesamaan karakteristik kontennya dengan preferensi pengguna di masa lalu. Sistem ini menganalisis fitur dari *item* seperti judul, deskripsi, atau kata kunci untuk mengukur tingkat kesamaan, dengan asumsi bahwa pengguna cenderung menyukai *item* yang serupa dengan yang pernah dipilih sebelumnya [5], [6]. Proses kerjanya dimulai dari ekstraksi fitur, misalnya menggunakan *TF-IDF* untuk merepresentasikan deskripsi produk dalam bentuk vektor numerik, kemudian membentuk profil pengguna dari interaksi sebelumnya, dilanjutkan dengan penghitungan kemiripan menggunakan metode seperti *cosine similarity*, dan akhirnya merekomendasikan *item* yang paling sesuai [5].

CBF memiliki beberapa keunggulan, seperti kemudahan implementasi karena hanya memerlukan atribut *item* tanpa data pengguna lain, kemampuan memberikan rekomendasi relevan meskipun basis pengguna kecil, serta personalisasi tinggi berdasarkan preferensi individu [6], [7]. Namun, metode ini juga memiliki keterbatasan, di antaranya masalah *cold start* untuk pengguna atau *item* baru, kecenderungan menciptakan *filter bubble* karena rekomendasi terlalu mirip dengan preferensi sebelumnya, dan ketergantungan pada kualitas fitur yang tersedia [5], [6].

Penerapan *CBF* telah dilakukan dalam berbagai studi. Ferdian *et al.* (2024) menggunakan *CBF* untuk merekomendasikan produk di *marketplace* berbasis kesamaan deskripsi dan kategori [5], sedangkan Kesuma & Iqbal (2020) menerapkannya dalam sistem rekomendasi penyedia jasa pernikahan dengan pendekatan *hybrid* untuk mengatasi data yang tidak lengkap [6]. Dinda Maristha *et al.* (2021) mengintegrasikan *CBF* dalam platform *e-commerce B2B* untuk merekomendasikan produk kesehatan, memanfaatkan fitur seperti komposisi dan indikasi medis yang dikaitkan melalui *graph database* [7].

Term Frequency–Inverse Document Frequency (*TF-IDF*) adalah metode yang digunakan untuk menilai seberapa penting suatu kata dalam sebuah dokumen relatif terhadap seluruh kumpulan dokumen atau korpus. *TF* mengukur frekuensi kemunculan kata dalam satu dokumen, sedangkan *IDF* menilai seberapa jarang kata tersebut muncul dalam keseluruhan dokumen. Dengan mengalikan keduanya, diperoleh nilai *TF-IDF* yang menyoroti kata-kata unik dan relevan dalam konteks tertentu [8]. Sebagai contoh, dalam kalimat "sistem rekomendasi menggunakan algoritma *machine learning*", kata “sistem” memiliki nilai *TF* 0,20. Jika kata “algoritma” hanya muncul di satu dari tiga dokumen, maka *IDF*-nya adalah log(3/1) = 0,477. Nilai *TF-IDF* sebesar 0,095 menunjukkan pentingnya kata tersebut dalam dokumen itu dibandingkan yang lain [9].

Keunggulan *TF-IDF* terletak pada kemampuannya menekankan kata kunci langka yang relevan dalam sebuah korpus, implementasi yang efisien, dan fleksibilitas untuk digabungkan dengan algoritma lain seperti *KNN* dan *Naïve Bayes* [8], [10]. Namun, kelemahannya adalah ketidakmampuan menangkap konteks semantik, tidak efektif dalam menangani sinonim atau ambiguitas makna, serta sensitivitas terhadap ukuran dan kualitas data [9].

Dalam implementasi penelitian, *TF-IDF* telah digunakan secara luas. Syuriadi & Astuti (2019) menerapkannya dalam klasifikasi *multi-label* teks hadis menggunakan *KNN* dengan *F1-score* 0,853, menunjukkan efektivitasnya dalam mengekstrak istilah penting seperti “zakat” atau “shalat” [8]. Kalokasari *et al.* (2017) menggunakannya untuk mengklasifikasikan surat keluar dan mencapai akurasi 89,58% dengan *Naïve Bayes* [10]. Sementara itu, Deolika *et al.* (2019) membuktikan bahwa *TF-IDF* masih kompetitif dibanding metode lain dalam klasifikasi dokumen, mencapai akurasi 98,67%, menjadikannya salah satu teknik dasar yang tetap relevan dalam *text mining* dan sistem rekomendasi [9].

Count Vectorization merupakan metode representasi teks ke dalam bentuk vektor numerik berdasarkan frekuensi kemunculan kata dalam dokumen. Teknik ini banyak digunakan dalam *NLP* karena mampu mengubah data teks menjadi input numerik yang dapat diproses oleh algoritma *machine learning* secara efisien [11]. Prosesnya mencakup tokenisasi untuk memisahkan kata, membangun *vocabulary* dari seluruh kata unik di korpus, serta menghitung frekuensi tiap kata di setiap dokumen. Hasilnya adalah matriks *document-term frequency* yang merepresentasikan setiap dokumen sebagai baris dan setiap kata unik sebagai kolom, dengan nilai dalam sel menunjukkan jumlah kemunculan kata tersebut [12].

Metode ini unggul karena sederhana, cepat diimplementasikan, dan efektif untuk dokumen pendek seperti *tweet* atau ulasan singkat [13]. Namun, ia memiliki keterbatasan dalam memahami makna atau konteks semantik, mengabaikan urutan kata, dan menghasilkan matriks yang cenderung *sparse* pada korpus besar, sehingga membutuhkan teknik lanjutan seperti reduksi dimensi [11].

Dalam implementasinya, *Count Vectorization* digunakan Dedy Sugiarto *et al.* (2022) untuk menganalisis sentimen terhadap kebijakan BLT minyak goreng melalui *tweet*, menghasilkan akurasi 0,72 dan *F1-score* 0,70 [13]. Rokhim (2017) memanfaatkannya untuk sistem pencarian skripsi berbasis *web* yang mengurutkan dokumen berdasarkan kemiripan teks terhadap kata kunci pencarian [12]. Sementara itu, Deo *et al.* (2024) membandingkannya dengan *TF-IDF* dan menunjukkan bahwa meskipun performa *Count Vectorization* sedikit lebih rendah (selisih 3–5% akurasi), waktu komputasinya lebih efisien hingga 40% [11]. Di samping itu, dalam forum akademik, *Count Vectorization* juga digunakan untuk mendeteksi topik utama diskusi mahasiswa berdasarkan frekuensi kata-kata dominan seperti "tugas" dan "*deadline*" [12].

HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) adalah pengembangan dari *DBSCAN* yang dirancang untuk menangani data dengan kepadatan bervariasi. Berbeda dari *DBSCAN* yang memerlukan nilai *eps* dan *minPts*, *HDBSCAN* membangun struktur hierarki berdasarkan jarak *mutual reachability* dan minimum *spanning tree* untuk mendeteksi *cluster* secara otomatis tanpa harus menentukan jumlahnya di awal [14]. Proses ini dilanjutkan dengan kondensasi hierarki dan ekstraksi *cluster* stabil menggunakan parameter *min_samples*, sekaligus memisahkan *data outlier*.

HDBSCAN unggul karena mampu mendeteksi *cluster* dengan densitas berbeda, tidak memerlukan input jumlah *cluster*, dan robust terhadap *outlier*. Namun, ia memiliki kompleksitas komputasi yang lebih tinggi serta tetap memerlukan parameter seperti *min_samples* dan pemahaman terhadap metrik jarak [15].

Dalam praktiknya, sebelum algoritme diterapkan, teks dokumen perlu diproses melalui tahapan pembersihan, tokenisasi, *stopword removal*, dan *stemming* untuk menghasilkan representasi yang lebih bersih. Dokumen kemudian direpresentasikan sebagai vektor melalui *TF-IDF*, *bag-of-words*, atau *embeddings*. *HDBSCAN* menganalisis distribusi kerapatan vektor tersebut, membentuk hierarki, dan mengekstrak *cluster* paling stabil, sedangkan data yang tidak cocok masuk ke dalam *cluster* utama akan dianggap sebagai *noise*.

Penelitian oleh Valles-Coral *et al.* (2022) membuktikan efektivitas *HDBSCAN* dalam mengelompokkan mahasiswa berdasarkan risiko *dropout* [14], sementara Neijenhuijs *et al*. (2021) berhasil mengelompokkan gejala pasien kanker secara lebih mendalam [15]. Dibandingkan *K-Means*, *HDBSCAN* tidak memerlukan nilai *k* dan menghasilkan *cluster* yang lebih valid berdasarkan metrik seperti *Silhouette Score* dan *Davies–Bouldin Index*, yang semakin sering digunakan sebagai standar evaluasi *internal clustering* [15], [16]. Meski demikian, evaluasi manual oleh pakar tetap penting untuk memastikan akurasi semantik dari hasil pengelompokan topik.

Cosine similarity adalah teknik pengukuran kemiripan teks yang didasarkan pada sudut antara dua vektor dalam ruang berdimensi banyak. Semakin kecil sudut antar vektor, semakin besar nilai *cosine similarity* yang dihasilkan, dengan rentang nilai antara 0 (tidak mirip) hingga 1 (identik). Nilai tersebut dihitung melalui *dot product* dibagi dengan hasil kali norma dari masing-masing vektor [17]. Metode ini sangat efektif dalam menangani representasi teks berdimensi tinggi seperti hasil dari *TF-IDF* atau *Count Vectorization* karena hanya fokus pada arah, bukan panjang dokumen [18].

Keunggulan *Cosine similarity* terletak pada kemampuannya membandingkan dokumen pendek dan panjang tanpa bias, serta efisiensinya dalam menangani *data sparse*, menjadikannya pilihan utama dalam sistem klasifikasi, deteksi plagiarisme, dan sistem rekomendasi [19], [20]. Studi oleh Rismayani *et al.* (2022) menunjukkan bahwa metode ini mampu mengklasifikasikan aspirasi publik ke lembaga legislatif dengan akurasi 92% [20], sedangkan Rio Feriangga Kurniawan (2022) menggunakannya untuk mengelompokkan konten berita di media sosial dengan keberhasilan tinggi dalam mengidentifikasi tema kecelakaan dan kriminal [17].

Selain itu, Sanjaya *et al.* (2023) menggabungkan *Cosine similarity* dengan basis data sinonim untuk meningkatkan kesesuaian semantik antar kalimat, meningkatkan rata-rata nilai kemiripan dari 0.78 menjadi 0.945 [18]. Dalam konteks deteksi plagiarisme, Lumbansiantar *et al.* (2023) memanfaatkan *TF-IDF* dan *Cosine similarity* untuk mengidentifikasi 15 kasus plagiat dari 100 jurnal *online* berdasarkan nilai kemiripan di atas 0.85 [19]. Konsistensinya dalam berbagai kasus membuat *cosine similarity* menjadi metode yang handal dan mudah diintegrasikan dengan pendekatan lain dalam pemrosesan teks.

K-Nearest Neighbor (*KNN*) adalah algoritma klasifikasi non-parametrik yang menentukan kelas suatu data berdasarkan kedekatannya dengan data lain dalam ruang fitur. Algoritma ini tidak memerlukan pelatihan model, melainkan menghitung kemiripan langsung antara data uji dan data pelatihan menggunakan metrik jarak seperti *Euclidean*, *Manhattan*, atau *cosine similarity* [21]. *KNN* bekerja dengan memilih *k* tetangga terdekat dari data baru, lalu mengklasifikasikannya berdasarkan mayoritas kelas yang dimiliki oleh tetangga tersebut [22].

Pemilihan nilai *k* memengaruhi akurasi, nilai kecil rentan terhadap *noise*, sementara nilai besar dapat menyebabkan generalisasi berlebih. Dalam konteks data teks, *cosine similarity* sering digunakan karena mampu mengukur arah kemiripan antar vektor *TF-IDF* atau *Count Vectorization* secara efektif [23].

KNN sangat berguna dalam evaluasi sistem rekomendasi berbasis konten karena dapat menilai seberapa baik vektorisasi teks dalam memetakan kemiripan semantik antar dokumen. Firdaus (2022) menggunakan *KNN* dengan *cosine similarity* untuk klasifikasi sentimen *Twitter* terkait *COVID-19* dan mendapatkan akurasi 85% pada *k*=15 [21]. Prasetya *et al.* (2019) menerapkan *KNN* untuk mendeteksi infeksi mulut berdasarkan sinyal suara, mencapai akurasi 87,5% menggunakan *correlation distance* [22]. Sementara itu, Fajri *et al.* (2020) menunjukkan bahwa *Manhattan distance* paling efektif dalam mengidentifikasi *radioisotop* dari data *spektroskopi* dengan akurasi 92% [24].

Dengan fleksibilitas dalam pemilihan metrik dan kemudahan implementasi, *KNN* menjadi alat evaluasi yang efisien untuk mengukur kualitas representasi fitur dalam berbagai domain klasifikasi, termasuk sistem rekomendasi berbasis teks.

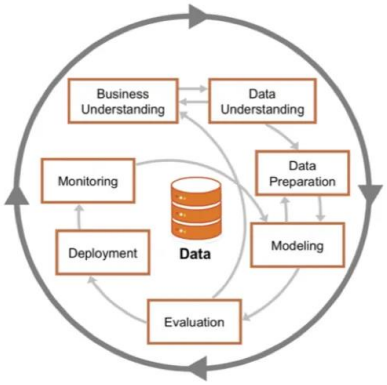
Confusion Matrix adalah alat evaluasi klasifikasi yang menggambarkan performa model melalui empat komponen utama: *True Positive* (*TP*), *False Positive* (*FP*), *True Negative* (*TN*), dan *False Negative* (*FN*). Matriks ini membantu mengidentifikasi jenis kesalahan prediksi, seperti tingginya nilai *FP* yang dapat menurunkan kepercayaan pengguna, atau *FN* yang berdampak serius pada sistem kritis seperti diagnosis medis [25], [26].

JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS)
E-ISSN : 2686-3359
Hal. 1807-1816
Vol. 7; No. 4
November 2025

Untuk menilai kinerja model secara kuantitatif, digunakan beberapa metrik evaluasi. *Accuracy* menunjukkan persentase prediksi yang benar dan cocok untuk *dataset* seimbang [10]. *Precision* mengukur ketepatan prediksi positif, sementara *recall* menilai kemampuan model dalam menangkap semua kasus positif, keduanya sangat penting dalam sistem yang sensitif terhadap kesalahan klasifikasi. *F1-score* menjadi metrik yang menyeimbangkan *precision* dan *recall*, terutama pada *dataset* tidak seimbang [13].

METODE

Penelitian ini menggunakan pendekatan *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*) sebagai kerangka metodologis dalam merancang sistem rekomendasi judul skripsi berbasis *content-based filtering*. Model *CRISP-DM* dipilih karena memiliki struktur yang sistematis dan fleksibel serta mampu mengintegrasikan tujuan bisnis dengan proses analisis data secara menyeluruh [27]. Dengan enam tahap utama yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Berikut merupakan uraian dari masing-masing tahapan yang dilakukan dalam penelitian ini.



Gambar 1. Tahapan CRISP-DM [28]

Business Understanding

Tahapan dimulai dari *business understanding* yang mengidentifikasi masalah pencarian judul skripsi di *repository* Universitas Jambi yang masih bersifat manual dan tidak efisien. Sebagai solusinya, dikembangkan sistem rekomendasi berbasis kemiripan konten menggunakan *TF-IDF* dan *Count Vectorization*, yang dievaluasi dengan algoritma *K-Nearest Neighbor* (*KNN*) dan metrik seperti *accuracy*, *precision*, *recall*, dan *F1-score*.

Data Understanding

Pada tahap *data understanding*, data dikumpulkan melalui *web scraping* dari *repository* UNJA, lalu dianalisis struktur dan kualitasnya, termasuk distribusi tahun, panjang teks, kategori jurusan/program studi, hingga deteksi bahasa menggunakan *langdetect*. Data yang tidak memiliki abstrak, duplikat, atau berbahasa non-Indonesia dihapus secara sistematis.

Data Preparation

Selanjutnya, dilakukan *data preparation* yang mencakup penggabungan kolom judul dan abstrak, pembersihan teks, tokenisasi, penghapusan *stopwords*, serta *stemming* untuk menghasilkan korpus teks yang siap digunakan.

Modeling

Tahap *modeling* dilakukan dengan membandingkan dua metode vektorisasi, *Count Vectorizer* dan *TF-IDF*. Keduanya digunakan untuk mengubah dokumen menjadi vektor numerik, yang kemudian dibandingkan dengan *Cosine Similarity* untuk menghitung tingkat kemiripan antar dokumen. Skor *similarity* digunakan untuk merekomendasikan judul-judul skripsi yang mirip berdasarkan kueri pengguna, dengan ambang batas minimum untuk menghindari rekomendasi yang tidak relevan.

Evaluation

Evaluasi dilakukan dalam dua tahap: *unsupervised* dan *supervised*. Pertama, label topik dibuat melalui klasterisasi menggunakan *HDBSCAN* dengan metrik *cosine* untuk mendeteksi kelompok tema yang relevan. Dokumen yang dikategorikan sebagai *noise* oleh *HDBSCAN* dihapus untuk menjaga akurasi. Label hasil klasterisasi kemudian digunakan dalam tahap *supervised learning* dengan algoritma *KNN*. Dua model *KNN* dilatih secara terpisah menggunakan representasi *Count Vectorizer* dan *TF-IDF*, lalu diuji menggunakan data uji sebesar 20% dari *dataset*. Evaluasi dilakukan dengan menghitung *Confusion Matrix*, *accuracy*, *precision*, *recall*, dan *F1-score* dengan pendekatan *weighted average* untuk mengakomodasi distribusi label yang tidak seimbang [10], [13].

Interpretasi hasil menekankan perbandingan *F1-score* dan *accuracy* dari kedua metode. Jika *TF-IDF* unggul, maka metode ini dianggap lebih efektif dalam menonjolkan kata khas suatu topik. Sebaliknya, jika *Count Vectorizer* memberikan hasil lebih baik, maka frekuensi kata sudah cukup dalam membedakan topik, dengan keuntungan komputasi yang lebih ringan. Selisih yang tidak signifikan (<1%) akan diuji secara statistik menggunakan *paired t-test* dengan $\alpha = 0,05$ untuk menentukan apakah kedua metode setara. Metode terbaik akan dinyatakan sebagai pemenang dan dijadikan mode default pada sistem.

JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS)
E-ISSN : 2686-3359
Hal. 1807-1816
Vol. 7; No. 4
November 2025

Deployment

Tahap akhir adalah *deployment* dalam bentuk aplikasi *demo* menggunakan *Streamlit* sebagai *proof of concept*. Aplikasi ini tidak diintegrasikan ke sistem *repository* UNJA, melainkan digunakan untuk validasi *internal* dan demonstrasi interaktif kepada pengguna dalam skenario terbatas.

HASIL DAN PEMBAHASAN

Business Understanding

Penelitian ini dilatarbelakangi oleh belum optimalnya sistem pencarian judul skripsi di *repository* Universitas Jambi, di mana mahasiswa masih kesulitan menemukan referensi yang relevan karena pencarian hanya mengandalkan kata kunci. Untuk mengatasi masalah tersebut, dikembangkan sistem rekomendasi berbasis *content-based filtering* yang mampu menyarankan judul skripsi serupa berdasarkan kemiripan isi konten, khususnya pada bagian judul dan abstrak.

Sistem ini menggunakan dua metode representasi teks, yaitu *TF-IDF* dan *Count Vectorization*, yang masing-masing dievaluasi untuk mengukur efektivitasnya dalam merepresentasikan konten dan menghasilkan rekomendasi yang relevan. Evaluasi dilakukan menggunakan algoritma *K-Nearest Neighbor* (*KNN*) untuk klasifikasi topik hasil rekomendasi, dan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score* dihitung melalui *Confusion Matrix*. Hasil evaluasi ini menjadi dasar untuk menilai sejauh mana sistem dapat memberikan rekomendasi yang sesuai dengan kebutuhan informasi mahasiswa.

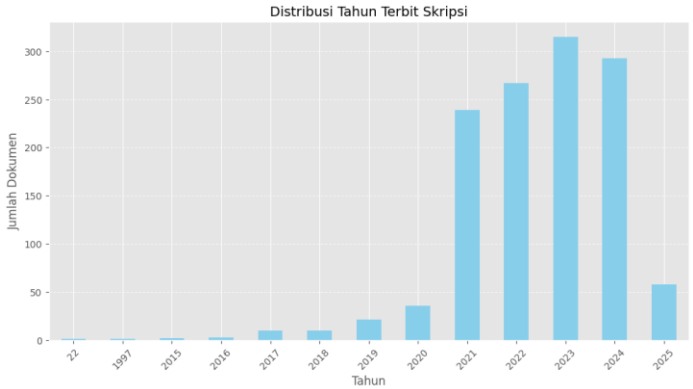
Data Understanding

Pada tahap ini, peneliti mengumpulkan sebanyak 1256 data skripsi melalui proses *web* scraping dari situs *repository.unja.ac.id*. *Dataset* yang diperoleh memuat enam atribut utama, yaitu *Link*, Info, Judul, Penulis, Tahun, dan Abstrak. Secara umum, seluruh kolom memiliki tipe data yang sesuai, namun ditemukan anomali pada kolom Tahun dengan nilai minimum 22 yang tidak logis dan perlu ditangani pada tahap pra-pemrosesan.

	Link	Info	Judul	Penulis	Tahun	Abstrak
0	https://repository.unja.ac.id/65405/	Analis Kimia	STUDI KELAYAKAN KUALITAS AIR MINUM PADA SAMPEL AIR SUMUR DI PERUMAHAN RAFIRA ASRI BERDASARKAN NILAI pH, NITRIT DAN KESADAHAN TOTAL	yanti, hasri	2024	Air memiliki unsur yang sangat penting bagi kehidupan manusia, tanpa air manusia akan mengalami kekurangan cairan, cairan dalam tubuh manusia seki...
1	https://repository.unja.ac.id/65421/	Analis Kimia	ANALISIS KADAR NITRAT (NO3\r\n-) DAN KLORIDA (Cl-) PADA SAMPEL AIR SUMUR DI \r\nPERUMAHAN RAFIRA ASRI SEBAGAI STUDI KELAYAKAN AIR BERSIH	ratih fortuna, miranda	2024	Air bersih merupakan salah satu kebutuhan pokok manusia yang \r\ndibutuhkan dalam kehidupan sehari-hari yang mana kebutuhannya semakin \r\nmeningk...
2	https://repository.unja.ac.id/65415/	Analis Kimia	ANALISIS KADAR KLORIDA (Cl-) SECARA ARGENTOMETRI DAN FLUORIDA (F-) MENGGUNAKAN METODE SPEKTROFOTOMETRI UV-Vis PADA AIR SUNGAI DI WILAYAH JAMBI	gustin, dea	2024	Air merupakan zat yang paling penting dalam kehidupan, sekitar tiga per empat bagian dari tubuh kita terdiri dari air. Air juga dipergunakan untuk...
3	https://repository.unja.ac.id/65417/	Analis Kimia	ANALISIS KADAR AMONIA (NH3) MENGGUNAKAN SPEKTROFOTOMETER UV-Vis DAN KADAR TOTAL DISSOLVED SOLIDS (TDS) SECARA GRAVIMETRI PADA AIR SUNGAI	Zuldekri, putri	2024	RINGKASAN\r\nAir merupakan salah satu sumber kehidupan makhluk hidup untuk kegiatan sehari-hari. Sungai adalah aliran air yang mengalir secara ter...
4	https://repository.unja.ac.id/65420/	Analis Kimia	ANALISIS KADAR SULFAT (SO42-) DAN TOTAL DISSOLVED SOLID (TDS) SEBAGAI PARAMETER UJI KELAYAKAN AIR BERSIH PADA SAMPEL AIR SUMUR	Sari, susnita	2024	Air merupakan bahan alam yang diperlukan untuk kehidupan manusia, hewan dan tanaman sebagai sumber energi serta berbagai keperluan lainnya. Air be...

Gambar 2. Sampel Dataset Repository UNJA

Hasil eksplorasi menunjukkan bahwa kolom *Abstrak* memiliki 49 nilai kosong (3,9%) dan terdapat 53 entri duplikat pada kolom Judul (4,22%). Dan distribusi Tahun menunjukkan puncak pada 2023, sementara anomali seperti "0022" akan ditangani di tahap *data cleaning*. Panjang rata-rata judul adalah 128 karakter, sedangkan abstrak bervariasi dengan rata-rata 2.044 karakter, mencerminkan kompleksitas konten yang cukup tinggi. Selain itu, analisis pada kolom Info mengungkapkan bahwa kategori "Sistem Informasi" merupakan penyumbang dokumen terbanyak (219 skripsi), yang perlu diantisipasi agar model tidak mengalami bias pada jurusan/program studi dominan. Deteksi bahasa menggunakan *langdetect* menunjukkan bahwa mayoritas abstrak (84,4%) berbahasa Indonesia, 11,7% berbahasa Inggris, dan sisanya tidak dapat diidentifikasi secara akurat. Informasi ini penting untuk memastikan konsistensi bahasa dalam pemrosesan teks. Secara keseluruhan, tingkat kelengkapan data mencapai 99,6%, dengan proporsi duplikat dan missing values masih dalam batas toleransi.



Gambar 3. Distribusi Tahun Terbit Skripsi

JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS)
E-ISSN : 2686-3359
Hal. 1807-1816
Vol. 7; No. 4
November 2025

Gambar 4. Top 10 Kategori dengan Dokumen Terbanyak

Data Preparation

Tahap *Data Preparation* dalam penelitian ini dilakukan untuk memastikan *dataset* hasil *web scraping* dari *repository* Universitas Jambi siap digunakan dalam pemodelan sistem rekomendasi. Proses dimulai dengan mengimpor berbagai pustaka penting seperti *pandas*, *NLTK*, dan *Sastrawi* untuk manipulasi data dan pemrosesan bahasa alami. Setelah *dataset* dimuat, dilakukan pembersihan awal terhadap nilai kosong dan duplikasi, terutama pada kolom Judul, yang menghasilkan 1180 data bersih dari duplikat.

Selanjutnya, diterapkan deteksi bahasa menggunakan kombinasi analisis probabilistik dan pencocokan kata kunci. Hanya abstrak berbahasa Indonesia yang dipertahankan, menghasilkan 882 data bersih yang konsisten secara linguistik. Kolom Judul dan Abstrak kemudian digabung ke dalam kolom *all_features* untuk menyatukan informasi penting. Teks kemudian dibersihkan dari simbol, angka, *URL*, dan tanda baca, lalu diubah menjadi huruf kecil melalui proses *case folding*.

Proses dilanjutkan dengan *tokenization* menggunakan *NLTK*, disusul penghapusan *stopwords* dari daftar standar dan tambahan manual. Kemudian, dilakukan *stemming* menggunakan *Sastrawi* untuk menyederhanakan bentuk kata ke bentuk dasarnya. Token hasil akhir digabung kembali menjadi teks utuh dalam kolom *processed_text*, dan data yang tidak valid dihapus.

Hasil akhirnya adalah 882 baris data yang siap digunakan untuk representasi fitur dan *modeling*, disimpan dalam dua file: satu untuk keperluan *modeling* dan satu sebagai cadangan lengkap. Tahapan ini memastikan bahwa seluruh data telah dibersihkan, dinormalisasi, dan disiapkan secara optimal untuk tahap selanjutnya.

Tabel 1. Hasil Pra-pemrosesan Teks

Teks Awal	**Teks Hasil Pra-pemrosesan**
Analisis User Experience Aplikasi Otentikasi Taspen Pada Nasabah PT. Taspen (Persero) Cabang Jambi Dengan Menggunakan Metode Enhanced Congnitive Walkthrough Taspen menghadirkan inovasi melalui aplikasi Otentikasi untuk mempermudah pensiunan dalam verifikasi kehadiran tanpa harus datang ke kantor setiap bulan. Namun, generasi baby boomer yang berusia â‰¥ 58 tahun ke atas menghadapi kesulitan dalam penggunaan aplikasi, terutama terkait user interface dan instruksi yang kurang jelas. Penelitian ini menganalisis user experience aplikasi menggunakan metode Enhanced Cognitive Walkthrough dengan lima responden pengguna baru dan merupakan nasabah PT.Taspen (persero) Cabang Jambi. Metode Enhanced Cognitive Walkthrough merupakan metode evaluasi yang berfokus pada model usability yaitu Learnability. Metode Enhanced Cognitive Walkthrough merupakan metode evaluasi user experience dimana responden diminta untuk mengerjakan tugas berbasis skenario yang telah dipersiapkan oleh peneliti. Hasil analisis mengindikasikan bahwa rata-rata tingkat permasalahan pada setiap matriks analisis berada pada tingkat rendah namun pada tugas yang sangat penting. Permasalahan paling banyak ditemukan yaitu pada task 6 (Ucapkan huruf A) yang juga memiliki nilai permasalahan tertinggi, diikuti oleh task 3 (Kedipkan mata) dan task 1 (Masukkan notas) yang menunjukkan tingkat permasalahan cukup signifikan. Dengan adanya evaluasi usability Otentikasi Taspen didapatkan 3 tampilan rekomendasi perbaikan dan 2 rekomendasi tambahan, dimana rekomendasi perbaikannya yaitu terdapat penambahan halaman petunjuk sebelum memasukkan notas, perbaikan instruksi kedipkan mata, perbaikan instruksi ucapkan huruf A, penambahan riwayat otentikasi pada halaman konfirmasi status otentikasi dan penambahan notifikasi otentikasi. Rekomendasi perbaikan dibuatkan berupa rancangan mockup sistem.	analisis user experience aplikasi otentikasi taspen nasabah taspen persero cabang jambi metode enhanced congnitive walkthrough taspen hadir inovasi aplikasi otentikasi mudah pensiun verifikasi hadir kantor generasi baby boomer usia hadap sulit guna aplikasi kait user interface instruksi teliti analis user experience aplikasi metode enhanced cognitive walkthrough responden guna nasabah pttaspen persero cabang jambi metode enhanced cognitive walkthrough metode evaluasi fokus model usability learnability metode enhanced cognitive walkthrough metode evaluasi user experience responden tugas bas skenario siap teliti hasil analisis indikasi ratarata tingkat masalah matriks analisis tingkat rendah tugas masalah temu task ucap huruf milik nilai masalah tinggi ikut task kedip mata task masuk notas tingkat masalah signifikan evaluasi usability otentikasi taspen dapat tampil rekomendasi baik rekomendasi tambah rekomendasi baik tambah halaman tunjuk masuk notas baik instruksi kedip mata baik instruksi ucap huruf tambah riwayat otentikasi halaman konfirmasi status otentikasi tambah notifikasi otentikasi rekomendasi baik buat rancang mockup sistem

Modeling

Tahap *Modeling* dimulai setelah proses *data preparation* menghasilkan korpus teks yang bersih dan konsisten. Sistem rekomendasi berbasis *content-based filtering* dibangun menggunakan dua metode vektorisasi teks, yaitu *Count Vectorizer* dan *TF-IDF*, serta pengukuran kemiripan antar dokumen menggunakan *Cosine Similarity*. *Dataset* yang telah diproses dimuat dan diperiksa kembali, kemudian kolom processed_*text* digunakan sebagai korpus utama.

Count Vectorizer mengubah teks menjadi representasi numerik berdasarkan frekuensi kemunculan kata, menghasilkan matriks berdimensi (882, 10268) dengan *sparsity* tinggi. Analisis menunjukkan kata "teliti", "hasil", dan "metode" paling sering muncul. Sebaliknya, *TF-IDF* memberikan bobot berdasarkan pentingnya istilah dalam konteks seluruh dokumen, dengan nilai rata-rata skor yang lebih kecil dan terdistribusi merata. Kata seperti "teliti", "air", dan "nilai" memiliki skor *TF-IDF* tertinggi, mencerminkan kontribusi semantiknya yang lebih relevan.

Tabel 2. Perbandingan Aspek Count Vectorizer dan TF-IDF

Aspek	**Count Vectorizer**	**TF-IDF**
Dimensi Matriks	(882, 10268)	(882, 10268)
Sparsity	99.23%	99.23%
Kata paling dominan	teliti, hasil, metode	teliti, air, nilai
Rata-rata Cosine Similarity	0.0889	0.0269

Perbandingan kedua metode menunjukkan bahwa *Count Vectorizer* cenderung memberikan skor kemiripan lebih tinggi, sementara *TF-IDF* lebih selektif dan konservatif. Hal ini terlihat dari nilai *Cosine Similarity* antar dokumen: *Count Vectorizer* memiliki nilai rata-rata 0.0889, sedangkan *TF-IDF* hanya 0.0269. Namun, korelasi antara keduanya sangat kuat (0.86), menandakan konsistensi relatif dalam pengukuran kemiripan meskipun skala nilainya berbeda.

Tabel 1. Perbandingan Hasil Rekomendasi

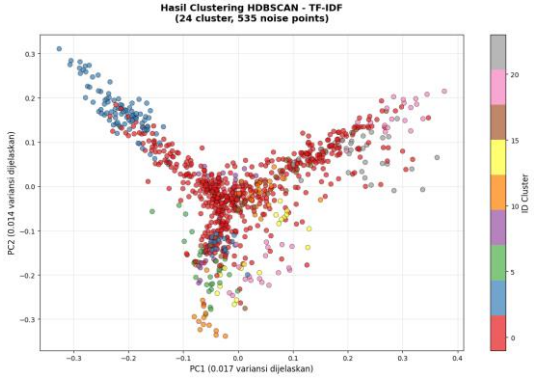
Dokumen Referensi	**Metode**	**Rekomendasi Dokumen (Index)**	**Similiarity Score**	**Topik Dokumen Terpilih**
Index 50 : Karakterisasi bakteri endofit daun sirih merah (Piper crocatum Ruiz & Pav) sebagai agen antibakteri terhadap Staphylococcus aureus dan Escherichia coli.	Count Vectorizer	30	0.4853	Efektivitas ekoenzim kulit nanas terhadap E. coli dan S. aureus
		44	0.4767	Karakterisasi bakteri endofitik daun bintaro (hasil antibiotik)
		48	0.4073	Uji aktivitas antibakteri ekstrak daun sungkai
		46	0.3863	Uji aktivitas antibakteri ekstrak daun Ageratum
		27	0.3674	Uji antibakteri ekstrak daun pegagan (C. asiatica)
	TF-IDF	30	0.3804	Efektivitas ekoenzim kulit nanas terhadap E. coli dan S. aureus
		44	0.3200	Karakterisasi bakteri endofitik daun bintaro (hasil antibiotik)
		48	0.2520	Uji aktivitas antibakteri ekstrak daun sungkai
		46	0.2326	Uji aktivitas antibakteri ekstrak daun Ageratum

Sistem rekomendasi diuji baik berdasarkan dokumen referensi. Hasilnya menunjukkan bahwa kedua metode mampu mengidentifikasi dokumen yang relevan secara tematik. *Count Vectorizer* unggul dalam menangkap kesamaan permukaan teks, sedangkan *TF-IDF* lebih baik dalam menyoroti istilah penting yang bersifat spesifik. Secara keseluruhan, kedua metode efektif untuk membangun sistem rekomendasi judul skripsi, dengan keunggulan masing-masing tergantung pada tujuan akhir: eksplorasi umum atau pencarian istilah spesifik.

Evaluation

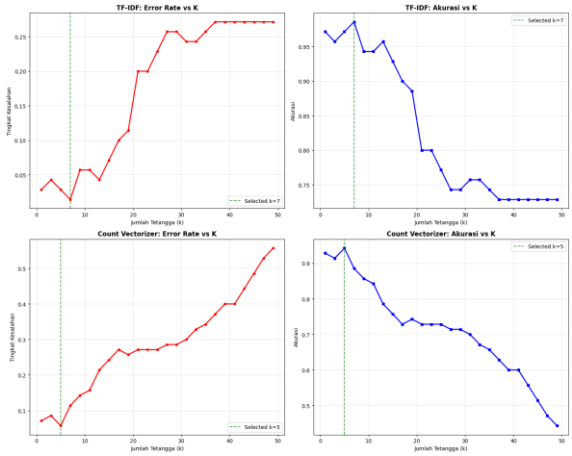
Tahap evaluasi dalam penelitian ini bertujuan menilai kinerja dua metode representasi teks, yaitu *TF-IDF* dan *Count Vectorizer*, dalam sistem rekomendasi berbasis kemiripan konten. Karena *dataset* abstrak skripsi tidak memiliki label topik eksplisit, pendekatan *semi-supervised* digunakan dengan membangkitkan label melalui klasterisasi *HDBSCAN*. Vektor *TF-IDF* dipilih sebagai input karena kemampuannya menyoroti kata-kata khas dalam dokumen. Proses klasterisasi menghasilkan 24 klaster dengan tingkat *noise* sebesar 60,66%. Meskipun nilai *Silhouette Score* hanya 0,1190, analisis isi tiap klaster menunjukkan kohesi tematik yang kuat, dari topik *AI* kedokteran, pertambangan, *biofuel*, hingga *geofisika*. Setelah menghapus *noise*, tersisa 347 dokumen yang digunakan untuk pelatihan model klasifikasi.

JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS)
E-ISSN : 2686-3359
Hal. 1807-1816
Vol. 7; No. 4
November 2025



Gambar 5. Hasil Clustering HDBSCAN

Data dibagi menjadi data latih dan uji, lalu dilakukan pelatihan *K-Nearest Neighbor*s (*KNN*) dengan variasi nilai *K*. Pada skema *TF-IDF*, akurasi terbaik (98,57%) dicapai pada K=7, sementara *Count Vectorizer* mencatat akurasi tertinggi (94,29%) pada K=5. Evaluasi menggunakan metrik Akurasi, Presisi, Recall, dan F1-*Score* menunjukkan dominasi *TF-IDF* di semua aspek. *TF-IDF* memberikan presisi 99,05% dan F1-*Score* 98,48%, jauh melampaui *Count Vectorizer* yang hanya mencapai F1-*Score* 94,15%.



Gambar 6. Visualisasi Optimasi Nilai K

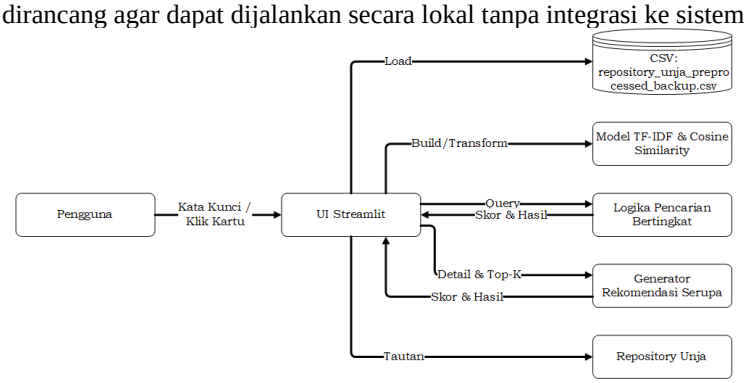
Visualisasi optimasi nilai K menegaskan stabilitas *TF-IDF* dalam berbagai skenario *K*, sementara *Count Vectorizer* cenderung menurun drastis setelah K>7. Perbandingan akhir menunjukkan bahwa model berbasis *TF-IDF + KNN* adalah pemenang eksperimen ini, karena mampu memberikan prediksi yang lebih akurat, stabil, dan relevan terhadap semantik dokumen. Pendekatan *TF-IDF* disimpulkan lebih unggul dan layak dijadikan standar dalam sistem rekomendasi judul skripsi berbasis teks.

Tabel 2. Hasil Evaluasi Model

Skema Representasi	**Akurasi**	**Presisi**	**Recall**	**F1-Score**
TF-IDF	0.9857	0.9905	0.9857	0.9848
Count Vectorizer	0.9429	0.9603	0.9429	0.9415

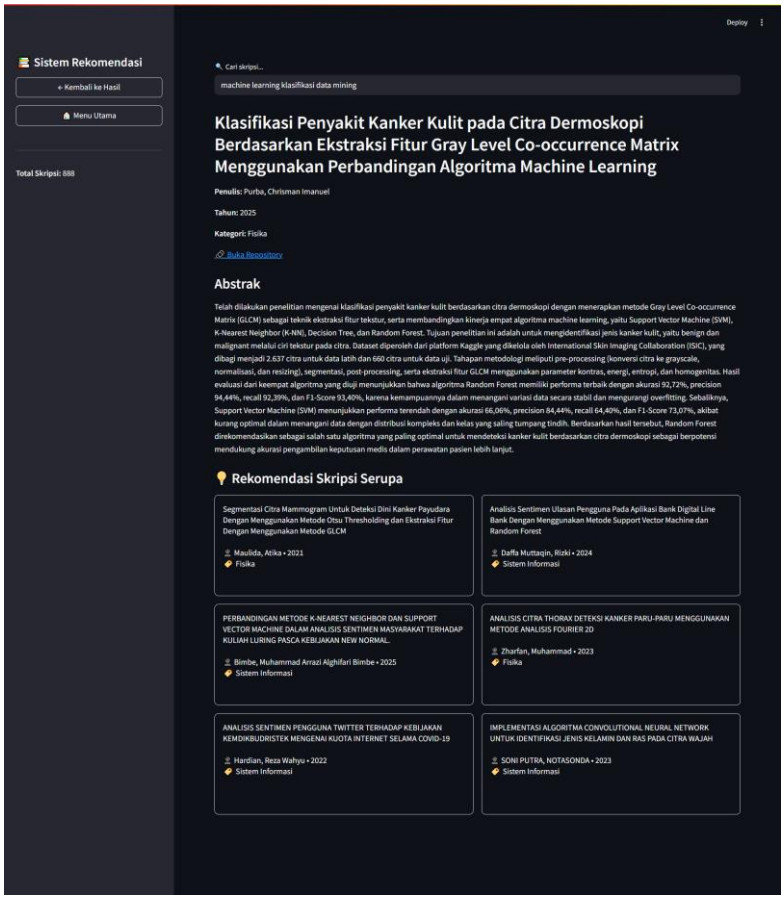
Deployment

Tahap *deployment* dalam penelitian ini dilakukan sebagai *proof-of-concept* untuk memvalidasi fungsionalitas sistem rekomendasi berbasis *TF-IDF* yang telah terbukti unggul dalam evaluasi sebelumnya. Aplikasi dibangun menggunakan *Streamlit* dan dirancang agar dapat dijalankan secara lokal tanpa integrasi ke sistem produksi.



Gambar 7. Diagram konteks/Level-0 DFD

Sistem ini mendemonstrasikan alur lengkap mulai dari pemuatan data, pencarian berdasarkan *query*, hingga rekomendasi judul serupa. Ketika aplikasi dijalankan, data dimuat dari file *repository_unja_preprocessed_backup.csv*, kemudian dibentuk representasi *TF-IDF* serta matriks *cosine similarity* antar dokumen. Pengguna dapat melakukan pencarian dengan memasukkan kata kunci, yang langsung diproses menjadi vektor dan dibandingkan menggunakan *Cosine Similarity*. Jika tidak ditemukan kecocokan melalui *TF-IDF*, sistem secara otomatis jatuh ke pencarian berbasis *substring* di seluruh kolom utama. Tampilan antarmuka dirancang sederhana dan intuitif. Pencarian menampilkan hasil dalam bentuk kartu dua kolom per halaman dengan sistem navigasi yang responsif. Setiap hasil bisa diklik untuk menampilkan halaman detail berisi informasi lengkap skripsi beserta abstraknya. Di bagian bawah, sistem juga menampilkan hingga enam rekomendasi skripsi lain yang paling mirip berdasarkan skor kemiripan tertinggi. Navigasi antar halaman dan reset *query* disediakan melalui sidebar untuk memudahkan pengguna.



Gambar 8. Tampilan UI Demo Detail Skripsi dan Rekomendasi Skripsi Serupa

KESIMPULAN DAN SARAN

Penelitian ini berhasil membandingkan efektivitas metode representasi teks *TF-IDF* dan *Count Vectorizer* dalam sistem rekomendasi judul skripsi berbasis *content-based filtering* di *repository* Universitas Jambi. Proses dilakukan secara sistematis mengikuti tahapan *CRISP-DM*, dari pengumpulan *data*, pembersihan, pembentukan label melalui klasterisasi *HDBSCAN*, hingga evaluasi menggunakan klasifikasi *K-Nearest Neighbor*s. Hasil menunjukkan bahwa *TF-IDF* membentuk klaster tematik yang lebih kuat dan mencapai akurasi tertinggi sebesar 98,57% pada K=7, jauh mengungguli *Count Vectorizer*. Dengan performa metrik yang konsisten lebih tinggi, *TF-IDF* terbukti mampu merepresentasikan konten secara lebih relevan dan presisi dalam konteks akademik.

DAFTAR PUSTAKA

[1] Muhammad Zaynurroyhan, Asriyanik, and Agung Pambudi, “Perbandingan TF-IDF dengan Count Vectorization Dalam Content-Based Filtering Rekomendasi Mobil Listrik,” *explorit,* vol. 15, no. 1, pp. 8–15, June 2023, doi: 10.35891/explorit.v15i1.3829.

[2] V. M. Hersianty, E. L. Amalia, D. Puspitasari, and D. W. Wibowo, “PENERAPAN ALGORITMA TF-IDF DAN COSINE SIMILARITY DALAM SISTEM REKOMENDASI LOWONGAN PEKERJAAN,” vol. 9, no. 1, 2025.

[3] D. S. Pradana, P. Prajoko, and G. P. Hartawan, “Perbandingan Algoritma Content-Based Filtering dan Collaborative Filtering dalam Rekomendasi Kegiatan Ekstrakurikuler Siswa,” *Progresif J. Ilmi. Kom,* vol. 18, no. 2, p. 151, July 2022, doi: 10.35889/progresif.v18i2.854.

[4] T. Ridwansyah, B. Subartini, and S. Sylviani, "Penerapan Metode Content-Based Filtering pada Sistem Rekomendasi," *Universitas Jambi*, vol. 4, no. 2, pp. 70–77, Apr. 2024, doi: 10.22437/msa.v4i2.32136.
[5] R. Ferdian, S. Achmady, and Z. Razi, "PENGEMBANGAN APLIKASI MARKETPLACE DENGAN PENERAPAN TEKNOLOGI MACHINE LEARNING BERBASIS WEB," vol. 3, no. 3, 2024.
[6] R. I. Kesuma and A. Iqbal, "Penerapan Content-Boosted Collaborative Filtering untuk Meningkatkan Kemampuan Sistem Rekomendasi Penyedia Jasa Acara Pernikahan," *FIFO*, vol. 12, no. 1, p. 112, May 2020, doi: 10.22441/fifo.2020.v12i1.009.
[7] M. D. Dinda Maristha, A. J. Santoso, and F. K. Sari Dewi, "Sistem Rekomendasi Pembelian Produk Kesehatan pada E-Commerce ABC berbasis Graph Database Amazon Neptune menggunakan Metode Hybrid Content-Collaborative Filtering," *JBI*, vol. 12, no. 2, pp. 88–97, Nov. 2021, doi: 10.24002/jbi.v12i2.4623.
[8] I. K. Syuriadi and W. Astuti, "Klasifikasi Teks Multi Label pada Hadis dalam Terjemahan Bahasa Indonesia Berdasarkan Anjuran, Larangan dan Informasi menggunakan TF-IDF dan KNN," 2019.
[9] A. Deolika, K. Kusrini, and E. T. Luthfi, "ANALISIS PEMBOBOTAN KATA PADA KLASIFIKASI TEXT MINING," *JurTI*, vol. 3, no. 2, p. 179, Dec. 2019, doi: 10.36294/jurti.v3i2.1077.
[10] D. H. Kalokasari, I. M. Shofi, and A. H. Setyaningrum, "IMPLEMENTASI ALGORITMA MULTINOMIAL NAIVE BAYES CLASSIFIER PADA SISTEM KLASIFIKASI SURAT KELUAR (Studi Kasus : DISKOMINFO Kabupaten Tangerang)," *J.Teknik Informatika*, vol. 10, no. 2, pp. 109–118, Oct. 2017, doi: 10.15408/jti.v10i2.6199.
[11] T. K. Deo, R. K. Deshmukh, and G. Sharma, "Comparative Study among Term Frequency-Inverse Document Frequency and Count Vectorizer towards K Nearest Neighbor and Decision Tree Classifiers for Text Dataset," *Nep. J. Multidisc. Res.*, vol. 7, no. 2, pp. 1–11, July 2024, doi: 10.3126/njmr.v7i2.68189.
[12] A. Rokhim, "IMPLEMENTASI METODE TERM FREQUENCY INVERSED DOCUMENT FREQUENCE (TF-IDF) DAN VECTOR SPACE MODEL PADA APLIKASI PEMBERKASAN SKRIPSI BERBASIS WEB," vol. 9, no. 1, 2017.
[13] Dedy Sugiarto, Ema Utami, and Ainul Yaqin, "Perbandingan Kinerja Model TF-IDF dan BOW untuk Klasifikasi Opini Publik Tentang Kebijakan BLT Minyak Goreng," *j. teknik industri*, vol. 12, no. 3, pp. 272–277, Dec. 2022, doi: 10.25105/jti.v12i3.15669.
[14] M. A. Valles-Coral *et al.*, "Density-Based Unsupervised Learning Algorithm to Categorize College Students into Dropout Risk Levels," *Data*, vol. 7, no. 11, p. 165, Nov. 2022, doi: 10.3390/data7110165.
[15] K. I. Neijenhuijs, C. F. W. Peeters, H. Van Weert, P. Cuijpers, and I. V. Leeuw, "Symptom clusters among cancer survivors: what can machine learning techniques tell us?," *BMC Med Res Methodol*, vol. 21, no. 1, p. 166, Dec. 2021, doi: 10.1186/s12874-021-01352-4.
[16] S. D'Amico *et al.*, "MOSAIC: An Artificial Intelligence–Based Framework for Multimodal Analysis, Classification, and Personalized Prognostic Assessment in Rare Cancers," *JCO Clin Cancer Inform*, no. 8, p. e2400008, June 2024, doi: 10.1200/CCI.24.00008.
[17] Rio Feriangga Kurniawan, "IMPLEMENTASI TEXT MINING MENGGUNAKAN METODE COSINE SIMILARITY UNTUK KLASIFIKASI KONTEN BERITA DI POSTINGAN GRUP FACEBOOK INFO LANTAS DAN KRIMINAL PASURUAN," *jami*, vol. 3, no. 1, pp. 9–17, June 2022, doi: 10.46510/jami.v3i1.41.
[18] A. Sanjaya, A. B. Setiawan, U. Mahdiyah, I. N. Farida, and A. R. Prasetyo, "Pengukuran Kemiripan Makna Menggunakan Cosine Similarity dan Basis Data Sinonim Kata," *JTIIK*, vol. 10, no. 4, pp. 747–752, Aug. 2023, doi: 10.25126/jtiik.20241046864.
[19] S. Lumbansiantar, S. Dwiasnati, and N. S. Fatonah, "Penerapan Metode Cosine Similarity Dalam Mendeteksi Plagiarisme Pada Jurnal," *FORMAT*, vol. 12, no. 2, p. 142, July 2023, doi: 10.22441/format.2023.v12.i2.007.
[20] R. Rismayani, H. Sy, T. Darwansyah, and I. Mansyur, "Implementasi Algoritma Text Mining dan Cosine Similarity untuk Desain Sistem Aspirasi Publik Berbasis Mobile," *Komputika*, vol. 11, no. 2, pp. 169–176, Aug. 2022, doi: 10.34010/komputika.v11i2.6501.
[21] A. Firdaus, "Aplikasi Algoritma K-Nearest Neighbor pada Analisis Sentimen Omicron Covid-19," *JRS*, pp. 85–92, Dec. 2022, doi: 10.29313/jrs.v2i2.1148.
[22] A. C. Prasetya, B. Hidayat, and R. Hartanto, "DETEKSI INFEKSI PADA RONGGA MULUT BERBASIS PEMROSESAN SINYAL WICARA DENGAN METODE DISCRETE COSINE TRANSFORM (DCT) DAN K NEAREST NEIGHBOR (KNN)," 2019.
[23] R. Kurnia, M. Asmita, R. Ihsan, I. Elfitri, and D. K. Hadi, "Perbandingan Metoda Klasifikasi K-Nearest Neighbor dan Support Vector Machine pada Pengenalan Benda Terhalang berbasis Kode Rantai," *ELKOMIKA*, vol. 12, no. 3, p. 823, July 2024, doi: 10.26760/elkomika.v12i3.823.
[24] M. S. Fajri, N. Septian, and E. Sanjaya, "Evaluasi Implementasi Algoritma Machine Learning K-Nearest Neighbors (kNN) pada Data Spektroskopi Gamma Resolusi Rendah," *Fiziya*, vol. 3, no. 1, pp. 9–14, Aug. 2020, doi: 10.15408/fiziya.v3i1.16180.
[25] R. R. Sani, Y. A. Pratiwi, S. Winarno, E. D. Udayanti, and F. A. Zami, "Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Hoax pada Berita Online Indonesia," vol. 13, no. 2, 2022.
[26] M. A. Afif, M. Ula, L. Rosnita, and R. Rizal, "Applying TF-IDF and K-NN for Clickbait Detection in Indonesian Online News Headlines," *Jo. Adv. Comp. Know. Algo*, vol. 1, no. 2, pp. 38–41, Apr. 2024, doi: 10.29103/jacka.v1i2.15810.
[27] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.
[28] SAP Community, "SAP Machine Learning: Approaching your Project," SAP Community. Accessed: Feb. 24, 2025. [Online]. Available: https://community.sap.com/t5/technology-blogs-by-sap/sap-machine-learning-approaching-your-project/ba-p/13359323