KETIK : Jurnal Informatika
ISSN: 3026-183X (Media Online)
Vol. 03, No. 06, Juli 2026 Hal. 267-274
Publisher: Faatuatua Media Karya
jurnal.faatuatua.com/index.php/KETIK
DOI: doi.org/10.70404/ketik.v3i06.693

Analisis Relevansi Judul Skripsi Teknik Informatika Universitas Halu Oleo (UHO) Menggunakan *TF-IDF*, *Cosine Similarity*, dan *Jaccard Similarity*

Ramlia Ramadani Sudin[1*], Elisa Wulan Sari[2], Adha Mashur Sajiah[3]

[1*]*Universitas Halu Oleo, Indonesia, e-mail: **ramadaniramlia@gmail.com***
[2]*Universitas Halu Oleo, Indonesia, e-mail: **elisawsari06@gmail.com***
[3]*Universitas Halu Oleo, Indonesia, e-mail: **adha.m.sajiah2@uho.ac.id***

Info Artikel

Diajukan: *29-06-2026*
Direvisi: *09-07-2026*
Diterima: *23-07-2026*
Diterbitkan: *Juli 2026*

Kata Kunci:
TF-IDF;
Cosine Similarity;
Jaccard Similarity;
K-Means clustering;
Judul Skripsi.

Keywords:
TF-IDF;
Cosine Similarity;
Jaccard Similarity;
K-Means clustering;
L-Thesis Titles.

Abstrak

Penelitian ini menganalisis kemiripan teks guna mengukur relevansi judul skripsi mahasiswa Program Studi Teknik Informatika Universitas Halu Oleo (UHO) menggunakan pendekatan information retrieval, yaitu TF-IDF, Cosine Similarity, dan Jaccard Similarity. Data berupa 620 judul skripsi UHO diperoleh dari dataset publik Kaggle, kemudian diproses melalui tahapan Case Folding, Cleaning, tokenisasi, penghapusan stopword (standar dan domain akademik), serta Stemming menggunakan PySastrawi. Tahap prapemrosesan berhasil mereduksi dimensi token sebesar 37,7%, meningkatkan efisiensi representasi dokumen. Representasi vektor dibangun menggunakan TF-IDF dengan kombinasi unigram dan bigram. Topik penelitian dikelompokkan menjadi delapan klaster menggunakan algoritma K-Means yang dikombinasikan dengan reduksi dimensi Latent Semantic Analysis (LSA). Evaluasi sistem temu kembali menggunakan 50 Query acak dengan ground truth berbasis klaster menunjukkan bahwa Jaccard Similarity secara konsisten mengungguli Cosine Similarity pada seluruh metrik pengujian (Precision@K, Recall@K, MAP, dan NDCG@K). Hasil ini mengindikasikan bahwa karakteristik judul skripsi yang ringkas menyebabkan sebaran bobot fitur TF-IDF rentan didominasi secara ekstrem oleh satu Term dominan tunggal yang memicu bias pada Cosine Similarity, sedangkan pendekatan kecocokan biner pada Jaccard Similarity terbukti lebih tangguh dan menghasilkan urutan peringkat dokumen yang jauh lebih akurat.

Abstract

This study analyzes Text similarity to measure the relevance of undergraduate thesis titles in the Department of Informatics Engineering at Universitas Halu Oleo (UHO) using information retrieval (IR) methods, specifically TF-IDF, Cosine Similarity, and Jaccard Similarity. A total of 620 UHO thesis titles obtained from a public Kaggle dataset were processed through Case Folding, Cleaning, Tokenization, stopword Removal (standard and academic domains), and Stemming using PySastrawi. The preprocessing stage successfully reduced token dimensions by 37.7%, improving document representation efficiency. TF-IDF vectors combining unigrams and bigrams were constructed for feature representation. Research topics were grouped into eight clusters using the K-Means algorithm combined with Latent Semantic Analysis (LSA) dimensionality reduction. Retrieval evaluation over 50 random queries using Cluster-Based Relevance proxy revealed that Jaccard Similarity consistently outperformed Cosine Similarity across all evaluation metrics (Precision@K, Recall@K, MAP, and NDCG@K). These results indicate that the concise nature of thesis titles makes TF-IDF feature weights susceptible to extreme dominance by a single Term, leading to bias in Cosine Similarity, whereas the binary token overlap approach in Jaccard Similarity proves more robust, yielding substantially more accurate document ranking recommendations.

1. PENDAHULUAN

Judul skripsi merupakan elemen fundamental dalam ranah akademik perguruan tinggi yang mencerminkan keseluruhan isi, ruang lingkup, tren kontribusi ilmiah, dan identitas kompetensi suatu penelitian. Di lingkungan Program Studi Teknik Informatika Universitas Halu Oleo (UHO), keragaman tema penelitian yang diajukan oleh mahasiswa terus mengalami peningkatan volume data yang signifikan dari tahun ke tahun. Relevansi antara judul skripsi yang diajukan dengan bidang keilmuan inti program studi menjadi aspek krusial untuk menjaga kualitas keluaran riset serta menghindari redundansi topik. Namun, evaluasi keselarasan tematik riset selama ini masih didominasi oleh metode penelaahan manual oleh tim komisi akademik atau dosen penguji. Proses penelaahan manual tersebut memerlukan alokasi waktu yang lama, tidak efisien untuk menangani ratusan berkas metadata skripsi, serta memiliki tingkat kerentanan yang tinggi terhadap subjektivitas penilai [1]. Oleh karena itu, penerapan pendekatan komputasional otomatis berbasis *information retrieval* (IR) dan pemrosesan bahasa alami atau *natural language processing* (NLP) menjadi sebuah kebutuhan mendesak guna mengukur tingkat kemiripan teks dokumen akademik secara objektif, efisien, dan terstruktur [2], [3].

Dalam bidang temu kembali informasi tekstual, representasi fitur numerik dokumen umumnya dibangun menggunakan metode *Term Frequency-Inverse Document Frequency* (*TF-IDF*) untuk mengukur bobot kepentingan suatu kata berdasarkan frekuensi kemunculannya di dalam dokumen relatif terhadap seluruh korpus [4]. Untuk mengukur derajat kedekatan atau kemiripan leksikal antara representasi vektor teks tersebut, terdapat dua metrik popular yang sering diimplementasikan, yaitu *Cosine Similarity* dan *Jaccard Similarity* [5], [6]. *Cosine Similarity* bekerja dengan cara menghitung nilai kosinus dari sudut yang terbentuk di antara dua vektor dalam ruang multidimensi, sehingga sangat sensitif terhadap pembobotan numerik yang dihasilkan oleh *TF-IDF* [7]. Di sisi lain, *Jaccard Similarity* menawarkan pendekatan biner yang lebih sederhana dengan mengukur rasio perbandingan antara irisan (*intersection*) dan gabungan (*union*) dari himpunan kosakata unik penyusun dokumen tanpa melibatkan pembobotan frekuensi kata [8]. Karakteristik data judul skripsi yang relatif pendek (memiliki keterbatasan jumlah kata) sering kali memicu fenomena sparsitas matriks yang tinggi, sehingga memengaruhi stabilitas dan kinerja dari kedua metrik kesamaan leksikal tersebut dalam menyajikan peringkat dokumen yang relevan [9], [10].

Sejumlah penelitian terdahulu telah berfokus pada pemanfaatan algoritma kemiripan teks untuk pengelolaan dokumen akademik. Hasanah et al. menerapkan algoritma Smith Wa*Term*an untuk mendeteksi kesamaan judul tugas akhir secara sekuensial karakter [1]. Sementara itu, Azmi memadukan pembobotan *TF-IDF* dan *Cosine Similarity* untuk mendeteksi *Persentase* plagiasi laporan skripsi [2]. Pemanfaatan arsitektur IR berbasis web juga dikembangkan oleh Amrulloh dan Adam untuk melakukan pencarian kemiripan judul tugas akhir mahasiswa [3]. Pada domain aplikasi yang berbeda, Abdullah dan Aribowo membangun purwarupa pengecekan judul skripsi Teknik Informatika [4], sedangkan Andriani dan Wibowo mengintegrasikan pembobotan kata untuk melakukan klasifikasi topik berbasis *content filtering* [5]. Evaluasi komparatif metrik jarak juga dilaporkan oleh Riyani et al. yang menyatakan bahwa *Cosine Similarity* memiliki performa yang baik dalam menangani ekstraksi fitur teks [6]. Dukungan komputasi ini diperkuat oleh Kurniadi et al. melalui otomasi pengarsipan dokumen universitas [7]. Kajian komparatif yang lebih luas dilakukan oleh Agustiawan untuk memetakan klaster artikel jurnal online [8], serta Kambey et al. yang memanfaatkan teknik *clustering* dokumen berbahasa Indonesia [9]. Terakhir, Amalia et al. menggunakan *Cosine Similarity* untuk melakukan penilaian otomatis pada jawaban teks pendek secara akurat [10].

Meskipun kajian literatur mengenai deteksi kemiripan naskah akademik telah banyak dieksplorasi, sebagian besar penelitian tersebut masih mengasumsikan penggunaan dokumen teks panjang secara umum dan belum mengevaluasi performa pencarian pada korpus spesifik judul skripsi pendek berbahasa Indonesia yang memiliki karakteristik *stopword* akademik yang unik [11]. Riset pembanding lain seperti analisis metadata oleh Syarif et al. lebih berfokus pada perbandingan *TF-IDF* dengan BM25 [12], sedangkan Nugraheni menguji perluasan himpunan melalui *Extended Jaccard* [13]. Penelitian struktural kemiripan dokumen oleh Nur et al. mengombinasikan *Jaccard* dengan Levenshtein Distance namun terbatas pada pemrosesan klaster bertahap atau blocking [14]. Selain itu, Rinartha dan Suryasa menguji pola *pattern matching* MySQL yang dibandingkan dengan *Jaccard* untuk saran kueri pencarian artikel [15]. Dari seluruh pemetaan literatur review tersebut, belum ada studi komparatif mendalam yang secara empiris membandingkan kinerja *Cosine Similarity* penuh berbasis *TF-IDF* unigram-bigram terhadap *Jaccard Similarity* biner pada korpus skripsi Teknik Informatika di lingkungan Universitas Halu Oleo.

Inovasi riset yang diajukan dalam penelitian ini terletak pada pengembangan *Pipeline* pemrosesan bahasa alami bahasa Indonesia yang mengintegrasikan ekstraksi fitur kueri kustom dengan perluasan *domain-specific stopword*s (seperti penghapusan *Term* adaptif akademik: "menggunakan", "berbasis", "implementasi", "rancang", "bangun") yang dikombinasikan dengan

pembentukan matriks unigram dan bigram guna meminimalkan dominasi kata struktural yang tidak informatif pada judul pendek. Selain itu, riset ini menerapkan inovasi pengujian sistem temu kembali informasi melalui evaluasi *unsupervised topic clustering* menggunakan kombinasi *Latent Semantic Analysis* (*LSA*) dan algoritma *K-Means* untuk membentuk proksi *ground truth* relevansi tematik yang objektif.

Berdasarkan permasalahan yang telah dijabarkan, penelitian ini bertujuan untuk melakukan analisis rekonstruksi ulang dan membandingkan performa kuantitatif antara metode *Cosine Similarity* berbasis bobot *TF-IDF* penuh dengan *Jaccard Similarity* berbasis himpunan token biner pada 620 data skripsi Teknik Informatika UHO. Evaluasi performa sistem temu kembali informasi diuji secara ketat menggunakan 50 kueri acak dan diukur melalui empat metrik standar internasional, yaitu *Precision at* K (*Precision@K*), *Recall at* K *(Recall@K*), *Mean Average Precision* (*MAP*), dan *Normalized Discounted Cumulative Gain* (*NDCG@K*). Hasil penelitian ini diharapkan dapat memberikan kontribusi teoretis mengenai perilaku metrik jarak pada teks singkat berbahasa Indonesia sekaligus memberikan rujukan aplikatif bagi pengelola program studi dalam *memonitoring*, mengevaluasi, serta mencegah duplikasi topik riset mahasiswa secara otomatis dan berbasis data ilmiah.

2. METODE PENELITIAN

Penelitian ini menerapkan pendekatan kuantitatif berbasis eksperimen komputasional untuk membangun sistem temu kembali informasi atau *information retrieval* (IR) pada dokumen teks pendek. Objek kajian utama dalam penelitian ini adalah keselarasan tematik pada kumpulan judul skripsi mahasiswa. Prosedur pelaksanaan eksperimen dirancang secara sistematis yang terbagi ke dalam lima tahapan utama, meliputi: isolasi sumber data sekunder, rekayasa draf *Pipeline* prapemrosesan teks, ekstraksi fitur *Term*, perhitungan metrik kemiripan teks, serta klasterisasi topik sebagai basis proksi pengujian sistem.

2.1 Sumber Data dan Isolasi Korpus

Data penelitian diperoleh secara sekunder dari repositori publik platform Kaggle bertajuk "*data-skripsi-its-uho*" yang diunggah oleh Aulia Adel [16]. Dataset mentah tersebut menggabungkan indeks metadata skripsi dari dua perguruan tinggi yang berbeda. Tahap isolasi data dilakukan secara terprogram melalui penyaringan string pada atribut Nomor Induk Mahasiswa (NIM). Dokumen milik Universitas Halu Oleo diidentifikasi secara spesifik melalui pencocokan pola karakter awal NIM yang menggunakan kode alfabetis "E1E", yang merepresentasikan identitas mahasiswa Program Studi Teknik Informatika UHO. Setelah proses eliminasi data asing dan pembersihan *missing values*, diperoleh korpus akhir sebanyak 620 baris dokumen judul skripsi berbahasa Indonesia yang siap digunakan sebagai basis data tekstual dalam eksperimen.

2.2 *Pipeline* Prapemrosesan Teks (*Text preprocessing*)

Karakteristik judul skripsi yang ringkas memerlukan penanganan prapemrosesan yang ketat untuk menghilangkan derau tekstual (*noise*) dan mereduksi sparsitas dimensi matriks fitur. *Pipeline* prapemrosesan teks yang dibangun dalam penelitian ini mengintegrasikan lima tahapan berurutan sebagai berikut:

1. *Case Folding*: Mengubah seluruh alfabet di dalam teks menjadi huruf kecil (*lowercase*) secara seragam untuk menghindari duplikasi token akibat perbedaan kapitalisasi.
2. *Cleaning*: Menghapus elemen non-alfabet seperti angka, tanda baca, simbol khusus, dan spasi ganda yang tidak esensial menggunakan ekspresi reguler (*regular expression*).
3. *Tokenization*: Memotong urutan string panjang dari judul skripsi menjadi kumpulan kata tunggal atau token individu.
4. *stopword Removal*: Mengeliminasi kata-kata umum yang memiliki frekuensi tinggi namun tidak memuat informasi pembeda topik tunggal. Proses ini memadukan daftar *stopword* standar bahasa Indonesia dari pustaka PySastrawi dengan *custom domain-specific stopwords* akademik yang disusun mandiri, meliputi kata: "menggunakan", "berbasis", "implementasi", "perancangan", "analisis", "studi", "kasus", "rancang", "bangun", "sistem", "aplikasi", "web", "mobile", "universitas", "halu", "oleo", "uho", "kendari", serta beberapa artefak kegagalan *Stemming* awal.
5. *Stemming*: Mengubah kata berimbuhan menjadi kata dasar menggunakan algoritma *stemmer* bawaan PySastrawi yang didesain khusus untuk morfologi bahasa Indonesia. Setelah itu, dilakukan penapisan akhir untuk membuang token yang memiliki panjang kurang dari 3 karakter.

2.3 Pembobotan Teks *TF-IDF* dan Ekstraksi Fitur

Vektorisasi teks dilakukan untuk mentransformasikan kumpulan token *space-separated* dari hasil prapemrosesan menjadi matriks numerik multidimensi. Metode pembobotan yang diterapkan adalah *Term Frequency-Inverse Document Frequency* (*TF-IDF*). Inovasi konfigurasi kelas pembentuk vektor menggunakan parameter *ngram_range=(1,2)* yang mengekstrak fitur kata tunggal (unigram) sekaligus kombinasi dua kata berturutan (bigram) untuk mempertahankan konteks frasa penting. Guna mereduksi fitur yang terlalu jarang atau terlalu dominan, diatur ambang batas frekuensi dokumen minimal *(min_df=2)* dan maksimal *(max_df=0.90)*. Selain itu, fungsi *sublinear_tf=True* diterapkan untuk mengaplikasikan transformasi logaritmik 1 + log(tf) pada frekuensi kemunculan *Term* guna meminimalkan bias dominasi kata berulang pada satu judul. Matriks bobot akhir kemudian dinormalisasi menggunakan fungsi norma L2.

2.4 Pengukuran Kemiripan Teks (*Similarity Computation*)

Matriks representasi fitur numerik yang terbentuk selanjutnya dihitung derajat kedekatan antar-dokumennya secara komparatif menggunakan dua metrik kesamaan teks, yaitu *Cosine Similarity* dan *Jaccard Similarity*.

1. *Cosine Similarity*: Mengukur orientasi arah sudut dari dua vektor *TF-IDF* dalam ruang dimensi tinggi. Nilai kesamaan kosinus dihitung berdasarkan perkalian titik (*dot product*) dari dua vektor dibagi dengan hasil perkalian magnitudo masing-masing vektor, dengan formula:

$$\cos(\mathrm{A},\mathrm{B}) = \frac{\mathrm{A}\cdot\mathrm{B}}{||\mathrm{A}||\,||\mathrm{B}||} \quad (1)$$

2. *Jaccard Similarity*: Mengukur kesamaan berdasarkan perbandingan leksikal biner dari himpunan token unik. Metrik ini mengabaikan bobot frekuensi numerik *TF-IDF* dan murni menghitung ukuran irisan token dibagi dengan ukuran gabungan token dari kedua dokumen judul, melalui persamaan:

$$\mathrm{J}(\mathrm{A},\mathrm{B}) = \frac{|\mathrm{A}\cap\mathrm{B}|}{|\mathrm{A}\cup\mathrm{B}|} \quad (2)$$

2.5 Klasterisasi Topik Menggunakan *K-Means* dan *LSA*

Untuk melakukan pengujian sistem temu kembali informasi tanpa adanya label relevansi manual dari pakar, penelitian ini membangun proksi relevansi berbasis klaster topik (*Cluster-Based Relevance proxy*). Tahapan klasterisasi dijalankan dengan mereduksi dimensi sparsitas matriks *TF-IDF* terlebih dahulu menggunakan teknik *Latent Semantic Analysis* (*LSA*) berbasis *Truncated Singular Value Decomposition* (*SVD*) sebanyak 50 komponen pembentuk varians. Vektor dimensi *LSA* tersebut kemudian dinormalisasi kembali menggunakan *L2-normalization* sebelum diumpankan ke dalam algoritma *K-Means clustering*. Penentuan jumlah kelompok optimal (K) dianalisis melalui metode *Elbow* dengan mengamati grafik penurunan nilai *Within-Cluster Sum of Squares* (*WCSS*) pada rentang K=4 hingga K=12. Dari hasil observasi grafik tersebut, ditemukan titik siku (*elbow*) yang paling optimal pada nilai K=8. Label kedekatan topik pada delapan klaster ini ditentukan berdasarkan rata-rata nilai *TF-IDF* tertinggi dari *Term* penyusun di setiap kelompok klaster.

2.6 Prosedur Pengujian dan Metrik Evaluasi Sistem

Sistem temu kembali informasi diuji kinerjanya secara ketat menggunakan 50 kueri dokumen skripsi yang diambil secara acak dari dataset korpus. Dua buah judul skripsi dianggap memiliki hubungan relevansi riil (*ground truth*) jika dan hanya jika kedua dokumen tersebut berada di dalam satu klaster topik *K-Means* yang sama. Berdasarkan skenario pencarian tersebut, sistem melakukan retrieval dan pemeringkatan rangking dokumen teratas menggunakan *Cosine Similarity* dan *Jaccard Similarity*. Perbandingan performa akurasi sistem diukur pada tiga variasi ambang batas jendela pencarian (K = 5, 10, dan 20) menggunakan empat metrik evaluasi standar *information retrieval* internasional, yaitu:

1. *Precision at* K (*Precision@K*): Mengukur proporsi ketepatan jumlah dokumen yang relevan dalam korpus hasil pencarian pada batas peringkat K teratas.
2. *Recall at* K *(Recall@K*): Mengukur rasio keberhasilan sistem dalam menemukan kembali dokumen relevan dari keseluruhan total dokumen yang seharusnya relevan di dalam klaster tersebut.
3. *Mean Average Precision* (*MAP*): Menghitung nilai rata-rata dari *Average Precision* (AP) di seluruh 50 kueri uji dengan mempertimbangkan presisi di setiap posisi peringkat dokumen relevan secara makro.

4. *Normalized Discounted Cumulative Gain* (*NDCG@K*): Mengukur efektivitas kualitas peringkat hasil pencarian dengan memberikan penalti bobot logaritmik yang semakin tinggi jika dokumen relevan berada di posisi peringkat bawah.

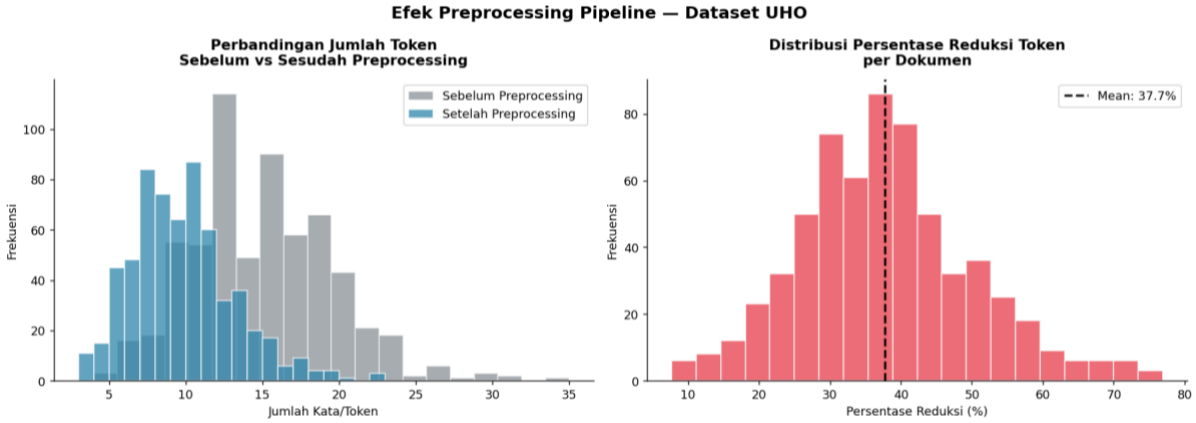
3. HASIL DAN ANALISIS

Pada bab ini dijabarkan analisis empiris hasil eksekusi program sistem temu kembali informasi menggunakan korpus judul skripsi Teknik Informatika Universitas Halu Oleo (UHO). Analisis meliputi statistik deskriptif efek prapemrosesan teks, pemetaan sebaran klaster topik menggunakan kombinasi *Latent Semantic Analysis* (*LSA*) dan K-Means, serta pengujian komparatif performa *Cosine Similarity* dan *Jaccard Similarity*.

3.2. Analisis Statistik Prapemrosesan Teks

Eksplorasi data awal terhadap 620 judul skripsi Teknik Informatika UHO menunjukkan karakteristik dokumen teks pendek dengan rata-rata panjang judul awal sebesar 14,99 kata (median 15 kata, standar deviasi 4,58). Proses prapemrosesan bertahap (*Case Folding*, *Cleaning*, *Tokenization*, *stopword Removal*, dan *Stemming*) terbukti secara signifikan mereduksi derau teks (*noise*).

Pasca-prapemrosesan, rata-rata jumlah token per judul berkurang menjadi 9,30 kata, yang mencerminkan *Persentase* reduksi rata-rata sebesar 37,7%. Hal ini mengonfirmasi bahwa sepertiga dari kosakata judul asli merupakan kata umum/struktural akademik (seperti "sistem", "menggunakan", "berbasis") yang tidak memiliki kekuatan diskriminatif dalam membedakan topik spesifik riset. Vektorisasi *TF-IDF* dengan rentang unigram dan bigram selanjutnya menghasilkan ruang fitur sebanyak 1.413 *Term* unik dengan tingkat kepadatan matriks (*matrix density)* sebesar 0,690%, yang mencerminkan sparsitas tinggi khas representasi teks singkat.



Gambar 1. Efek Prapemrosesan *Pipeline* terhadap Reduksi Jumlah Token

3.3. Analisis Sebaran Klaster Topik (*K-Means* + *LSA*)

Berdasarkan metode *Elbow* pada nilai *Within-Cluster Sum of Squares* (*WCSS*), jumlah kelompok dokumen yang paling optimal ditetapkan pada K=8. Reduksi dimensi melalui *Truncated SVD* (*LSA*) dengan 50 komponen utama menghasilkan varians terjelaskan sebesar 29,7%. Distribusi 620 dokumen skripsi ke dalam 8 klaster topik memperlihatkan pola sebaran yang tidak merata, sebagaimana disajikan secara terstruktur pada Tabel 1.

Tabel 1. Distribusi dan Kata Kunci Representatif Klaster Topik Skripsi UHO

Klaster	Top 5 *Term*	Jumlah Judul	*Persentase*
0	internet, *internet of things*, *things*, *monitoring*, *IoT*	48	7,7%
1	putus, dukung, dukung putus, terima, tunjang putus	57	9,2%
2	android, cari, kamus, *search*, *smartphone android*	60	9,7%
3	data, kota, jual, sms, *clustering*	130	21,0%
4	*fuzzy*, pakar, sakit, logika, logika *fuzzy*	52	8,4%
5	*server*, aman, *file*, jaring, *client*	81	13,1%
6	informatika, deteksi, citra, identifikasi, *game*	132	21,3%
7	informasi, kota, informasi geografis, geografis, jual	60	9,7%

Data pada Tabel 1 mengindikasikan bahwa tren penelitian mahasiswa Teknik Informatika UHO didominasi oleh dua arah utama. Pertama adalah pengembangan sistem informasi berbasis pengolahan data (Klaster 3 mencakup 21,0%). Kedua adalah domain pengolahan citra digital, deteksi objek, dan pengembangan *game* (Klaster 6 mencakup 21,3%). Sebaliknya, topik-topik mutakhir berskala global seperti *internet of things* (*IoT*) dan sistem *monitoring* terdistribusi masih sangat minim dieksplorasi (Klaster 0 hanya mencakup 7,7%).

Gambar 2. Visualisasi Sebaran Klaster Topik Skripsi UHO via Proyeksi *LSA* 2D

3.4. Evaluasi Perbandingan Kinerja Sistem

Pengujian performa pencarian sistem temu kembali dilakukan menggunakan 50 kueri acak dengan mengacu pada *Cluster-Based Relevance* proxy. Ringkasan hasil evaluasi kuantitatif akurasi perbandingan antara *Cosine Similarity* berbasis bobot *TF-IDF* penuh dengan *Jaccard Similarity* berbasis biner disajikan secara mendalam pada Tabel 2.

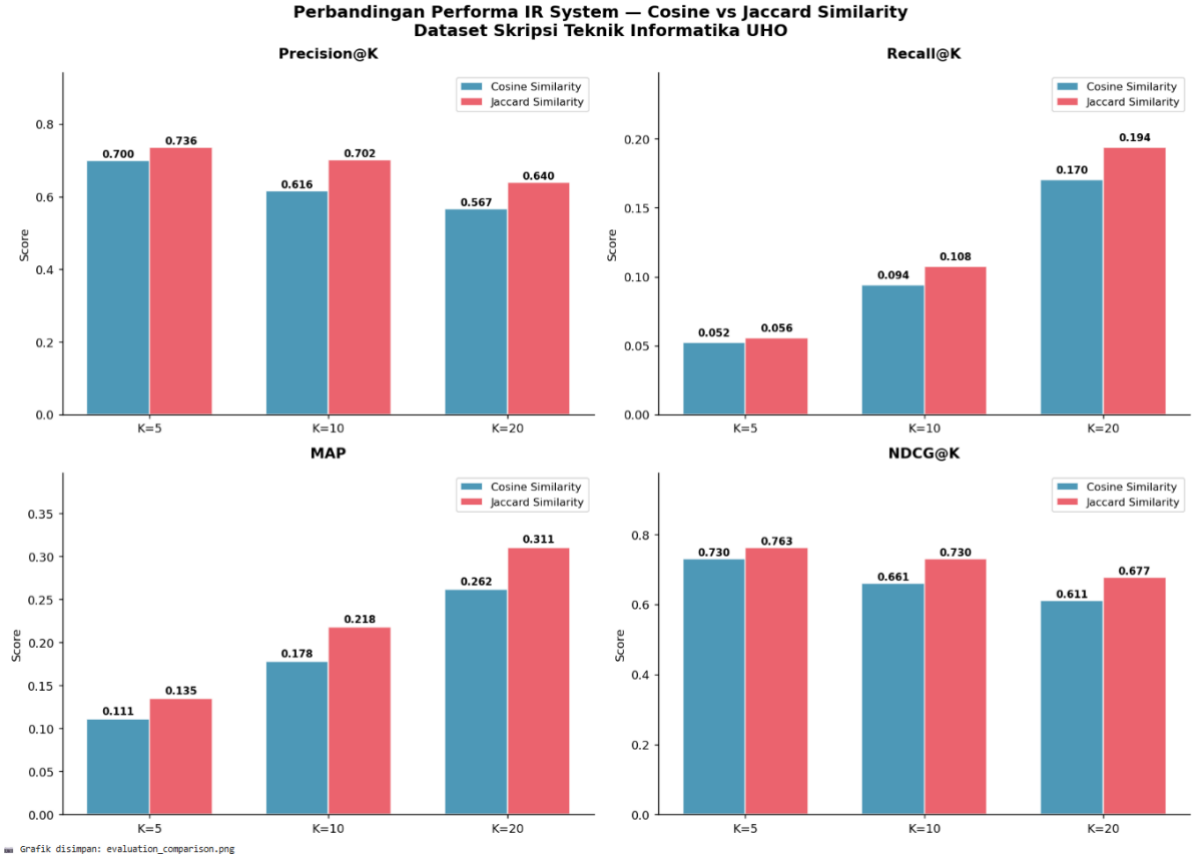
Tabel 2. Hasil Evaluasi Sistem Temu Kembali (50 *Query*, *Cluster-Based Relevance*)

Metode	K	*Precision@K*	*Recall@K*	*MAP*	*NDCG@K*
Cosine Similarity	5	0,7000	0,0523	0,1107	0,7296
Cosine Similarity	10	0,6160	0,0941	0,1776	0,6611
Cosine Similarity	20	0,5670	0,1704	0,2617	0,6109
Jaccard Similarity	5	0,7360	0,0557	0,1345	0,7632
Jaccard Similarity	10	0,7020	0,1077	0,2179	0,7299
Jaccard Similarity	20	0,6400	0,1939	0,3105	0,6769

Berdasarkan data empiris pada Tabel 2, *Jaccard Similarity* secara konsisten mengungguli *Cosine Similarity* pada seluruh metrik evaluasi (*Precision@K*, *Recall@K*, *MAP*, dan *NDCG@K*) di setiap tingkatan nilai K yang diuji. Sebagai contoh, pada batas jendela pencarian utama K=10, *Jaccard Similarity* memperoleh skor *Precision*@10 sebesar 0,7020, *MAP* sebesar 0,2179, dan *NDCG*@10 sebesar 0,7299. Nilai tersebut melampaui capaian *Cosine Similarity* yang masing-masing hanya memperoleh skor *Precision*@10 sebesar 0,6160, *MAP* sebesar 0,1776, dan *NDCG*@10 sebesar 0,6611.

Analisis statistik lebih lanjut menunjukkan bahwa meskipun kedua metrik ini memiliki nilai korelasi Pearson yang sangat kuat sebesar 0,9311 pada pengujian sampel 500 pasangan dokumen

(mencerminkan konsistensi penangkapan pola kemiripan secara umum), keduanya memiliki sensitivitas yang sangat kontras ketika dihadapkan pada karakteristik teks pendek.



Gambar 2. Grafik Batang Perbandingan Nilai Metrik Evaluasi Kinerja IR

Fenomena keunggulan *Jaccard Similarity* ini terjadi karena karakteristik judul skripsi yang sangat pendek cenderung memicu sebaran bobot nilai *TF-IDF* yang tidak merata (*flat distribution*) dan rentan didominasi secara ekstrem oleh satu atau dua *Term* bernilai tinggi. Sebagai contoh, pada kasus duplikasi semu antara judul "Perancangan *game* Maze 3D Menggunakan Algoritma Backtracking" dan "Penerapan Algoritma A* *Pathfinding* Dalam Pembentukan *Artificial Intelligence Enemy* Pada *game*", nilai kesamaan Cosine tercatat bernilai sangat tinggi akibat kesamaan *Term* dominan "game". Dominasi fitur tunggal ini menyebabkan *Cosine Similarity* kurang stabil dan memicu bias pencarian dokumen yang salah.

Sebaliknya, *jaccard similarity* yang berbasis kecocokan himpunan token biner terbukti jauh lebih kokoh *(robust)* terhadap gangguan bias dominasi bobot kata berulang tersebut. Pendekatan biner *Jaccard* lebih sensitif dalam menangkap irisan kosakata riil secara menyeluruh pada dokumen ringkas, sehingga menghasilkan urutan peringkat hasil pencarian (*retrieval ranking*) yang secara substansi tematik jauh lebih akurat dan relevan bagi kebutuhan pemantauan duplikasi dokumen akademik skripsi.

4. KESIMPULAN

Penelitian ini berhasil menganalisis kemiripan teks guna mengukur relevansi judul skripsi mahasiswa Program Studi Teknik Informatika Universitas Halu Oleo menggunakan pendekatan sistem temu kembali informasi. Implementasi komponen *Pipeline* pemrosesan bahasa alami bahasa Indonesia yang mengintegrasikan penapisan domain-specific *stopword*s akademik terbukti efektif dalam mereduksi dimensi kosakata tak informatif pada dokumen teks pendek sebesar 37,7%. Berdasarkan pengujian evaluasi temu kembali informasi yang diukur secara objektif melalui proksi relevansi berbasis klaster topik, disimpulkan bahwa metode *Jaccard Similarity* biner secara konsisten memiliki performa akurasi pemeringkatan rangking yang lebih baik dan lebih stabil dibandingkan dengan metode *Cosine Similarity* berbasis nilai pembobotan *TF-IDF* penuh. Karakteristik judul skripsi yang ringkas menyebabkan sebaran bobot fitur *TF-IDF* rentan didominasi secara ekstrem oleh satu *Term* dominan tunggal, yang memicu bias orientasi arah sudut pada perhitungan *Cosine Similarity*. Sebaliknya, pendekatan kecocokan elemen biner pada *Jaccard Similarity* terbukti lebih tangguh terhadap gangguan sebaran bobot tersebut sehingga mampu menyajikan rekomendasi urutan dokumen yang secara

substansi jauh lebih relevan. Secara keseluruhan, riset ini memberikan kontribusi rujukan aplikatif bagi pihak pengelola program studi dalam membangun sistem *monitoring* tren topik penelitian sekaligus mendeteksi potensi duplikasi judul tugas akhir mahasiswa secara otomatis berbasis data ilmiah. Pengembangan lebih lanjut disarankan untuk menerapkan representasi makna tingkat tinggi seperti *semantic word embeddings* bahasa Indonesia untuk menangkap kedekatan makna kontekstual yang berada di luar kecocokan leksikal teks singkat.

UCAPAN TERIMAKASIH

Penulis menyampaikan terima kasih yang sebesar-besarnya kepada Program Studi Teknik Informatika Universitas Halu Oleo atas dukungan atmosfer akademik dan bimbingan yang telah diberikan selama masa perkuliahan dan penyusunan penelitian ini. Apresiasi yang tinggi juga ditujukan kepada kontributor platform repositori publik Kaggle atas penyediaan akses dataset metadata akademik yang menjadi basis data utama dalam riset ini.

REFERENSI

[1] U. N. Hasanah, R. Satra, and F. Umar, "Deteksi Kemiripan Judul Skripsi Menggunakan Algoritma Smith Wa*Term*an," *Buletin Sistem Informasi dan Teknologi (BUSITI)*, vol. 1, no. 1, pp. 56–65, 2020.

[2] M. Azmi, "Analisis Tingkat Plagiasi Dokumen Skripsi Dengan Metode *Cosine Similarity* Dan Pembobotan *TF-IDF*," *Jurnal Teknik, Teknologi Informasi dan Multimedia*, vol. 2, no. 2, pp. 90–95, 2022.

[3] A. Amrulloh and I. F. Adam, "Sistem Pencarian Similaritas Judul Tugas Akhir Menggunakan Metode *TF-IDF*," *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi*, vol. 7, no. 2, pp. 74–82, 2021.

[4] I. Abdullah and E. Aribowo, "Rancang Bangun Aplikasi Pengecekan Kemiripan Judul Skripsi Dengan Metode *Cosine Similarity* (Studi Kasus : Program Studi Teknik Informatika UAD)," *Jurnal Informatika*, vol. 6, no. 2, pp. 43–52, 2018.

[5] N. Andriani and A. Wibowo, "Implementasi *Text* Mining Klasifikasi Topik Tugas Akhir Mahasiswa Teknik Informatika Menggunakan Pembobotan *TF-IDF* dan Metode *Cosine Similarity* Berbasis Web," *Prosiding Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, pp. 130–137, September 2021.

[6] A. Riyani, M. Zidny, and A. Burhanuddin, "Penerapan *Cosine Similarity* dan Pembobotan *TF-IDF* untuk Mendeteksi Kemiripan Dokumen," *Jurnal Komputasi*, vol. 2, no. 1, pp. 23–27, 2019.

[7] D. Kurniadi, S. F. C. Haviana, and A. Novianto, "Implementasi Algoritma *Cosine Similarity* pada sistem arsip dokumen di Universitas Islam Sultan Agung," *Jurnal Transformatika*, vol. 17, no. 2, p. 124, 2020.

[8] A. Agustiawan, "Analisis Similarity/Kemiripan Artikel Jurnal Online Terbitan Tahun 2019-2020 Di ISI Yogyakarta," *ABDI PUSTAKA: Jurnal Perpustakaan dan Kearsipan*, vol. 2, no. 1, pp. 29–43, 2022.

[9] G. E. I. Kambey, et al., "Penerapan *clustering* pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia," *Jurnal Sains dan Teknologi*, vol. 15, no. 2, pp. 75–82, 2020.

[10] E. L. Amalia, A. J. Jumadi, I. A. Mashudi, and D. W. Wibowo, "Analisis Metode *Cosine Similarity* Pada Aplikasi Ujian Online Otomatis (Studi Kasus JTI POLINEMA)," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 8, no. 2, p. 343, 2021.

[11] S. M. D. A. Jan, "Integrasi *TF-IDF* dan Algoritma *Cosine Similarity* untuk Deteksi Tingkat Kemiripan Judul Penelitian Teknik Informatika," Skripsi, Program Studi Teknik Informatika, Universitas Ichsan Gorontalo, Gorontalo, 2023.

[12] M. Syarif, R. Sa'adah, M. R. A. Listi, and R. Manisha, "Analisis Perbandingan Algoritma BM25 dan *TF-IDF* untuk Temu Kembali Metadata Jurnal Indonesia pada Temujurnal.com," *Jurnal ICT: Information Communication & Technology*, vol. 25, no. 2, pp. 173–180, Desember 2025.

[13] M. Nugraheni, "Perbandingan *Jaccard Similarity* dengan Extended *Jaccard Similarity* pada Penalaran Berbasis Kasus," *Jurnal PINTER: Pendidikan Informatika dan Sains*, vol. 4, no. 2, pp. 45–52, Desember 2020.

[14] M. R. Nur, G. S. Buana, and N. A. Rakhmawati, "Analisis Komparatif Pengukuran Kemiripan Artikel Ilmiah menggunakan *Jaccard* dan Levenshtein serta Blocking," *Jurnal Teknik Informatika dan Sistem Informasi (JUTISI)*, vol. 9, no. 2, pp. 1541–1553, Agustus 2023.

[15] K. Rinartha and W. Suryasa, "Comparative study for better result on *Query* suggestion of article *search*ing with MySQL pattern matching and *Jaccard Similarity*," in *Proceedings of the 5th International Conference on Cyber and IT Service Management (CITSM)*, Denpasar, 2017, pp. 1–4.

[16] A. Adel, "data-skripsi-its-uho," *Kaggle*, 2024. (*Available*: https://www.kaggle.com/datasets/auliaadel/data-skripsi-its-uho).