# Analisis Preprocessing Pada Similarity Judul Skripsi Menggunakan TF-IDF Dan Cosine Similarity

**Eviana Tjatur Putri[1], Mohamad Ardi[2]**

[1]Sistem Informasi, STMIK PPKIA Tarakanita Rahmawati, Tarakan, Kalimantan Utara
[2]Teknik Informatika, STMIK PPKIA Tarakanita Rahmawati, Tarakan, Kalimantan Utara
Email: [1]eviana@ppkia.ac.id, [2]mohamadardi@ppkia.ac.id

**Abstrak**

Kemiripan judul skripsi dapat menyebabkan pengulangan topik penelitian dan mengurangi variasi penelitian mahasiswa. Penelitian ini bertujuan menganalisis pengaruh tahapan preprocessing terhadap nilai similarity judul skripsi menggunakan metode TF-IDF dan Cosine Similarity. Dataset penelitian terdiri atas 480 judul skripsi bidang teknologi informasi yang digunakan sebagai data pembanding dalam proses perhitungan similarity. Pengujian sistem dilakukan menggunakan 50 judul uji yang berada di luar dataset, terdiri atas 25 judul skripsi dari periode sebelumnya dan 25 judul usulan terbaru pada periode Genap 2025. Tahapan preprocessing yang dianalisis meliputi raw text, cleaning, stopword removal, dan stemming. Hasil penelitian menunjukkan bahwa setiap tahapan preprocessing menghasilkan variasi nilai similarity. Pada kelompok judul lama, nilai similarity rata-rata berturut-turut sebesar 36,89%, 37,07%, 36,61%, dan 38,28%, sedangkan pada kelompok judul baru sebesar 39,18%, 39,44%, 39,07%, dan 39,62%. Tahap stemming menghasilkan nilai similarity rata-rata tertinggi dibandingkan tahapan preprocessing lainnya pada kedua kelompok data uji. Namun, peningkatan yang diperoleh relatif kecil sehingga menunjukkan bahwa pengaruh preprocessing terhadap nilai similarity pada dataset penelitian cenderung terbatas. Penelitian ini juga menghasilkan sistem berbasis web yang dapat digunakan untuk membantu proses evaluasi pengajuan judul skripsi secara lebih cepat dan objektif.

**Kata Kunci:** Preprocessing, Text Mining, TF-IDF, Cosine Similarity, Judul Skripsi.

# *Analyzing the Impact of Text Preprocessing on Thesis Title Similarity Using TF-IDF and Cosine Similarity*

***Abstract***

*The similarity of undergraduate thesis titles may lead to the repetition of research topics and reduce the diversity of student research. This study aims to analyze the effect of preprocessing stages on thesis title similarity using the TF-IDF and Cosine Similarity methods. The dataset consists of 480 information technology thesis titles used as reference data for similarity computation. System evaluation was conducted using 50 testing titles outside the reference dataset, comprising 25 thesis titles from previous academic periods and 25 newly proposed titles submitted in the 2025 even semester. The preprocessing stages evaluated include raw text, cleaning, stopword removal, and stemming. The results indicate that each preprocessing stage produces variations in similarity values. For the previous thesis title group, the average similarity values were 36.89%, 37.07%, 36.61%, and 38.28%, respectively, while the corresponding values for the newly proposed title group were 39.18%, 39.44%, 39.07%, and 39.62%. Among the evaluated preprocessing stages, stemming produced the highest average similarity values for both testing groups. However, the improvement was relatively small, indicating that the effect of preprocessing on similarity values was limited for the dataset used in this study. In addition, this research developed a web-based system that can support a faster and more objective evaluation of undergraduate thesis title submissions.*

***Keywords:*** *Preprocessing, Text Mining, TF-IDF, Cosine Similarity, Thesis Title*

## I. PENDAHULUAN

Perkembangan teknologi informasi memberikan dampak besar terhadap pemanfaatan data digital dalam lingkungan pendidikan tinggi. Salah satu bentuk pemanfaatannya adalah penerapan text mining untuk membantu pengolahan dokumen akademik secara otomatis. Permasalahan yang sering muncul

pada proses akademik adalah adanya kemiripan judul skripsi yang diajukan mahasiswa. Kemiripan judul tersebut dapat menyebabkan pengulangan topik penelitian, berkurangnya variasi penelitian, hingga potensi kesamaan ide penelitian. Oleh karena itu, diperlukan suatu mekanisme yang mampu membantu proses identifikasi kemiripan judul skripsi secara lebih cepat, objektif, dan efisien [1],[2].

Pada banyak perguruan tinggi, proses pemeriksaan kemiripan judul masih dilakukan secara manual dengan membaca dan membandingkan judul satu per satu. Proses tersebut menjadi kurang efisien ketika jumlah data skripsi terus bertambah setiap periode. Selain membutuhkan waktu yang relatif lama, proses manual juga berpotensi menimbulkan subjektivitas dalam menentukan tingkat kemiripan antar judul. Kondisi ini mendorong perlunya penerapan metode komputasi yang mampu melakukan pengukuran similarity dokumen secara otomatis [3],[4].

Pengukuran kemiripan dokumen berbasis teks umumnya dilakukan melalui kombinasi metode pembobotan dan metode perhitungan kedekatan vektor. Pada penelitian ini digunakan TF-IDF untuk menghasilkan bobot setiap term berdasarkan karakteristik kemunculannya dalam dokumen, sedangkan Cosine Similarity digunakan untuk menghitung tingkat kesamaan antar dokumen dari representasi vektor yang terbentuk. Kombinasi kedua metode tersebut telah banyak digunakan dalam bidang text mining dan information retrieval karena mampu memberikan hasil pengukuran kemiripan yang cukup baik pada berbagai jenis dokumen, termasuk judul skripsi dan dokumen akademik lainnya [5],[6],[7].

Dalam proses text mining, tahapan preprocessing memiliki peranan penting karena mempengaruhi kualitas representasi teks yang digunakan dalam proses pembobotan dan pengukuran similarity. Tahapan preprocessing umumnya meliputi case folding, cleaning, tokenizing, stopword removal, dan stemming. Penerapan tahapan tersebut bertujuan untuk menghasilkan data yang lebih terstruktur sehingga proses pengolahan teks dapat dilakukan secara lebih optimal [3],[8]. Namun demikian, pengaruh masing-masing tahapan preprocessing terhadap hasil similarity tidak selalu sama. Karakteristik data yang berbeda dapat menghasilkan perubahan nilai similarity yang berbeda pula setelah preprocessing diterapkan.

Beberapa penelitian penerapan metode TF-IDF dan Cosine Similarity telah dilakukan pada deteksi kemiripan dokumen maupun melakukan proses temu kembali informasi [2],[3]. Akan tetapi, sebagian besar penelitian masih berfokus pada implementasi metode dan hasil similarity yang diperoleh. Penelitian yang membahas pengaruh setiap tahapan preprocessing terhadap nilai similarity, khususnya pada data judul skripsi bidang teknologi informasi, masih relatif terbatas. Selain itu, belum banyak penelitian yang melakukan evaluasi pengaruh preprocessing menggunakan sejumlah data uji yang mewakili karakteristik judul skripsi dari periode yang berbeda.

Sebagai upaya untuk mendukung proses evaluasi usulan judul skripsi, penelitian ini mengembangkan aplikasi berbasis web yang menerapkan metode TF-IDF dan Cosine Similarity dalam pengukuran tingkat kemiripan judul. Penelitian tidak hanya berfokus pada implementasi sistem, tetapi juga mengkaji dampak tahapan preprocessing terhadap hasil pengukuran similarity yang diperoleh. Data yang digunakan terdiri atas 480 judul skripsi bidang teknologi informasi dari periode Ganjil 2022 hingga Ganjil 2025 sebagai data pembanding. Pengujian dilakukan terhadap 50 judul, yang mencakup 25 judul dari periode sebelumnya dan 25 judul usulan terbaru pada periode Genap 2025. Hasil penelitian diharapkan dapat memberikan gambaran mengenai kontribusi preprocessing dalam proses pengukuran kemiripan serta mendukung evaluasi usulan judul skripsi secara lebih efektif dan objektif.

Beberapa penelitian sebelumnya telah menerapkan metode TF-IDF dan Cosine Similarity dalam pengukuran kemiripan dokumen. Penelitian oleh Wahyuni dkk. menerapkan algoritma Cosine Similarity dan pembobotan TF-IDF pada sistem klasifikasi dokumen skripsi dan menunjukkan tingkat ketepatan klasifikasi yang cukup tinggi [9].

Penelitian lain dilakukan oleh Nasrullah yang mengintegrasikan TF-IDF dan Cosine Similarity untuk mendeteksi kemiripan tugas akhir dengan memanfaatkan tahapan preprocessing berupa case folding, tokenizing, stopword removal, dan stemming [10].

Selain itu, Andriani menerapkan text mining, TF-IDF, dan Cosine Similarity untuk klasifikasi topik tugas akhir mahasiswa berbasis web menggunakan data skripsi perguruan tinggi [11]. Penelitian Lim dkk. menerapkan TF-IDF dan Cosine Similarity pada proses text summarization ulasan aplikasi Mobile JKN. Hasil penelitian menunjukkan bahwa kombinasi kedua metode tersebut mampu mengidentifikasi informasi penting pada dokumen teks secara efektif [7].

Berdasarkan penelitian terdahulu tersebut, sebagian besar penelitian masih berfokus pada implementasi metode similarity dan pengembangan sistem deteksi kemiripan dokumen. Analisis mengenai pengaruh masing-masing tahapan preprocessing terhadap nilai similarity masih relatif terbatas, khususnya pada data judul skripsi bidang teknologi informasi. Selain itu, evaluasi pengaruh preprocessing menggunakan kelompok data uji yang berbeda karakteristik dan periode penelitian juga masih jarang dilakukan.

Penelitian yang dilakukan tidak hanya mengimplementasikan metode TF-IDF dan Cosine Similarity pada sistem deteksi kemiripan judul skripsi, tetapi juga menganalisis pengaruh setiap tahapan preprocessing melalui evaluasi terhadap dua kelompok data uji yang terdiri dari judul skripsi periode sebelumnya dan judul pengajuan terbaru.

## II. METODOLOGI PENELITIAN

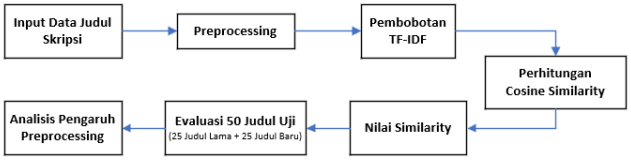
Text mining merupakan proses ekstraksi informasi dari data teks tidak terstruktur melalui tahapan pengolahan dan analisis tertentu [8]. Melalui text mining, kumpulan dokumen dapat dianalisis untuk menemukan pola tertentu, hubungan antar kata, maupun tingkat kemiripan antar dokumen. Penerapan text mining saat ini banyak digunakan pada sistem pencarian informasi, analisis sentimen, klasifikasi dokumen, dan pengukuran similarity [12],[13].

Pada prosesnya, data teks yang awalnya tidak terstruktur akan diubah menjadi bentuk yang lebih mudah diproses oleh komputer. Tahapan tersebut biasanya melibatkan preprocessing, pembobotan term, dan proses pengukuran similarity [10],[14].

## A. Tahapan Penelitian

Penelitian ini dilakukan melalui beberapa tahapan yang meliputi pengumpulan data, preprocessing text, pembobotan TF-IDF, penghitungan similarity menggunakan Cosine Similarity, serta analisis pengaruh preprocessing terhadap hasil similarity judul skripsi. Tahapan penelitian dilakukan secara bertahap untuk memperoleh hasil pengukuran similarity yang optimal

Adapun alur penelitian ditunjukkan pada Gambar 1.




Gambar 1. Alur Penelitian

.

## B. Dataset Penelitian

Penelitian ini menggunakan 480 data judul skripsi bidang teknologi informasi yang diperoleh dari database akademik sebagai data pembanding dalam proses pengukuran similarity. Setiap data memiliki atribut berupa judul skripsi, periode, dan dosen pembimbing.

Untuk mengevaluasi pengaruh preprocessing terhadap hasil similarity, penelitian ini juga menggunakan 50 judul uji yang berada di luar dataset tersebut, terdiri dari 25 judul skripsi periode sebelumnya dan 25 judul pengajuan terbaru periode genap 2025. Data uji tersebut digunakan untuk membandingkan nilai similarity yang dihasilkan pada setiap tahapan preprocessing, yaitu raw text, cleaning, stopword removal, dan stemming.

## C. Preprocessing Text

Preprocessing merupakan tahap awal dalam text mining yang bertujuan mempersiapkan data teks agar lebih siap digunakan pada proses analisis. Tahapan ini dilakukan untuk meningkatkan kualitas representasi data dengan menghilangkan unsur-unsur yang tidak relevan sehingga hasil pengukuran kemiripan dapat menjadi lebih akurat [10],[12].

Pada penelitian ini, preprocessing terdiri atas beberapa tahapan. Case folding digunakan untuk menyeragamkan seluruh karakter menjadi huruf kecil (lowercase) sehingga tidak terdapat perbedaan representasi antara huruf kapital dan huruf nonkapital. Tahap cleaning dilakukan dengan menghilangkan karakter yang tidak diperlukan, seperti angka, simbol, dan tanda baca. Selanjutnya, tokenizing digunakan untuk memecah teks menjadi sejumlah token atau kata yang dapat diproses lebih lanjut. Stopword removal diterapkan untuk menghapus kata-kata umum yang memiliki kontribusi rendah terhadap proses pengukuran kemiripan. Tahap terakhir adalah stemming, yaitu proses mengubah kata ke bentuk dasarnya agar kata-kata yang memiliki makna serupa dapat direpresentasikan sebagai term yang sama [10],[14].

Seluruh data judul skripsi pada penelitian ini diproses melalui tahapan preprocessing sebelum dilakukan pembobotan menggunakan TF-IDF dan perhitungan similarity dengan Cosine Similarity. Tahapan yang diterapkan meliputi case folding, cleaning, tokenizing, stopword removal, dan stemming. Proses tersebut dilakukan baik pada data judul skripsi yang tersimpan dalam database maupun pada judul baru yang dimasukkan oleh pengguna melalui sistem. Untuk stopword removal digunakan daftar stopword Bahasa Indonesia yang disimpan pada database, sedangkan proses stemming memanfaatkan pustaka Sastrawi yang diimplementasikan menggunakan bahasa pemrograman Python.

## D. D. Pembobotan TF-IDF

TF-IDF (Term Frequency–Inverse Document Frequency) merupakan metode yang digunakan untuk memberikan bobot pada setiap term berdasarkan tingkat kepentingannya dalam suatu dokumen. Penentuan bobot dilakukan dengan mempertimbangkan frekuensi kemunculan term pada dokumen tertentu serta distribusinya pada seluruh koleksi dokumen [7],[9],[10].

Konsep dasar TF-IDF adalah memberikan bobot yang lebih tinggi pada kata yang sering muncul dalam suatu dokumen, namun memiliki frekuensi kemunculan yang rendah pada dokumen lainnya. Dengan demikian, term yang dianggap lebih representatif terhadap isi dokumen akan memperoleh nilai bobot yang lebih besar. Oleh karena itu, metode TF-IDF banyak dimanfaatkan pada berbagai penelitian text mining dan information retrieval.

Persamaan TF-IDF ditunjukkan sebagai berikut:

$$TF - IDF(t,d) = TF(t,d)\ x\ IDF(t) \quad (1)$$

Pada Persamaan (1), TF(t,d) menunjukkan jumlah kemunculan term t pada dokumen d, sedangkan IDF(t) merepresentasikan nilai inverse document frequency yang digunakan untuk mengukur tingkat keunikan term dalam kumpulan dokumen.

Setelah tahapan preprocessing selesai dilakukan, setiap judul skripsi direpresentasikan ke dalam bentuk vektor numerik menggunakan metode TF-IDF. Proses ini bertujuan mengubah data teks menjadi nilai numerik yang dapat digunakan dalam perhitungan similarity. Pada penelitian ini, pembobotan TF-IDF diimplementasikan menggunakan pustaka scikit-learn pada bahasa pemrograman Python sesuai dengan Persamaan (1).

## E. E. Pengukuran Similarity

Cosine Similarity merupakan metode yang digunakan untuk mengukur tingkat kemiripan antara dua dokumen yang telah direpresentasikan dalam bentuk vektor. Perhitungan dilakukan dengan membandingkan sudut yang terbentuk antara dua vektor tersebut sehingga dapat diketahui tingkat kedekatan isi dokumen [10],[14].

Nilai Cosine Similarity berada pada rentang 0 hingga 1. Semakin mendekati nilai 1, semakin tinggi tingkat kemiripan antara dokumen yang dibandingkan. Sebaliknya, nilai yang

mendekati 0 menunjukkan bahwa kedua dokumen memiliki tingkat kesamaan yang rendah. Karena mampu membandingkan dokumen berdasarkan representasi vektor, metode ini banyak digunakan dalam text mining dan information retrieval, termasuk pada data teks pendek seperti judul skripsi dan abstrak penelitian.

Persamaan Cosine Similarity ditunjukkan sebagai berikut:

$$\cos(\theta) = \frac{A-B}{\|A\|\|B\|} \quad (2)$$

Pada Persamaan (2), A dan B merepresentasikan vektor dokumen yang dibandingkan, sedangkan ‖A‖ dan ‖B‖ menunjukkan panjang masing-masing vektor.

Dalam penelitian ini, Cosine Similarity digunakan untuk menghitung tingkat kemiripan antara judul yang diinputkan pengguna dengan seluruh data judul skripsi yang tersimpan pada database. Nilai similarity yang diperoleh kemudian digunakan sebagai dasar untuk menentukan urutan judul yang paling relevan terhadap judul masukan.

Hasil pengukuran ditampilkan dalam bentuk persentase kemiripan dan diurutkan berdasarkan nilai similarity tertinggi sehingga pengguna dapat dengan mudah mengidentifikasi judul skripsi yang memiliki tingkat kedekatan paling tinggi dengan judul yang diajukan..

### F. Implementasi Sistem

Penelitian ini diimplementasikan dalam bentuk sistem berbasis web menggunakan framework Flask pada bahasa pemrograman Python. Sistem mampu menerima input judul skripsi dari pengguna, melakukan preprocessing, menghitung similarity menggunakan TF-IDF dan Cosine Similarity, serta menampilkan 10 judul skripsi dengan similarity tertinggi.

Selain itu, sistem juga melakukan analisis pengaruh preprocessing terhadap hasil similarity melalui perbandingan similarity pada setiap tahapan preprocessing yang digunakan.

### G. Evaluasi Pengujian

Evaluasi dilakukan untuk menganalisis pengaruh setiap tahapan preprocessing terhadap nilai similarity yang dihasilkan oleh metode TF-IDF dan Cosine Similarity.

Setiap judul uji diproses melalui empat skenario pengolahan teks, yaitu raw text, cleaning, stopword removal, dan stemming. Pada setiap skenario, sistem menghitung nilai similarity tertinggi terhadap 480 data judul skripsi yang terdapat pada database.

Hasil similarity dari seluruh data uji kemudian dihitung nilai rata-ratanya untuk mengetahui pengaruh masing-masing tahapan preprocessing terhadap proses pengukuran kemiripan dokumen. Nilai rata-rata similarity digunakan sebagai dasar analisis dalam menentukan tahapan preprocessing yang memberikan hasil lebih baik.

## III. HASIL DAN PEMBAHASAN

### A. Implementasi Sistem

Penelitian ini menghasilkan sistem berbasis web yang digunakan untuk mendeteksi kemiripan judul skripsi menggunakan metode TF-IDF dan Cosine Similarity. Sistem dibangun menggunakan framework Flask pada bahasa pemrograman Python dan terhubung dengan database yang berisi 480 data judul skripsi bidang teknologi informasi.

Sistem menerima input berupa judul skripsi baru dari pengguna, kemudian melakukan tahapan preprocessing yang meliputi case folding, cleaning, tokenizing, stopword removal, dan stemming. Setelah proses preprocessing selesai dilakukan, sistem melakukan pembobotan TF-IDF dan penghitungan similarity menggunakan Cosine Similarity.

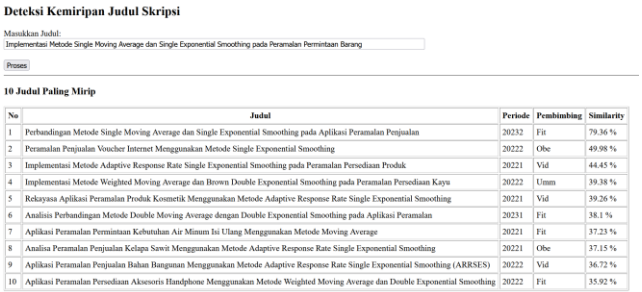
Hasil similarity ditampilkan dalam bentuk persentase kemiripan dan diurutkan berdasarkan nilai similarity tertinggi. Selain itu, sistem juga menampilkan 10 judul skripsi dengan similarity tertinggi beserta informasi periode dan dosen pembimbing.


Gambar 2. Halaman Input Deteksi Kemiripan Judul Skripsi

### B. Hasil Pengujian Similarity

Pengujian dilakukan dengan memasukkan judul skripsi baru ke dalam sistem untuk menghitung tingkat similarity terhadap 480 data judul skripsi yang terdapat pada database. Proses similarity dilakukan menggunakan metode TF-IDF dan Cosine Similarity setelah melalui tahapan preprocessing text.

Gambar 3 menunjukkan contoh hasil similarity yang dihasilkan sistem.




Gambar 3. Hasil Deteksi Kemiripan Judul Skripsi

Berdasarkan hasil pengujian, sistem mampu menampilkan 10 judul skripsi dengan tingkat kemiripan tertinggi terhadap judul yang diinputkan pengguna. Informasi yang ditampilkan meliputi judul skripsi, periode, dosen pembimbing, serta persentase similarity yang dihitung menggunakan metode TF-IDF dan Cosine Similarity.

**Tabel 1.** Lima Judul dengan Similarity Tertinggi

| No | Judul | Similarity (%) |
|---|---|---|
| 1 | Perbandingan Metode Single Moving Average dan Single Exponential Smoothing pada Aplikasi Peramalan Penjualan | 79.36 |
| 2 | Peramalan Penjualan Voucher Internet Menggunakan Metode Single Exponential Smoothing | 49.98 |

| | | |
|---|---|---|
| 3 | Implementasi Metode Adaptive Response Rate Single Exponential Smoothing pada Peramalan Persediaan Produk | 44.45 |
| 4 | Implementasi Metode Weighted Moving Average dan Brown Double Exponential Smoothing pada Peramalan Persediaan Kayu | 39.38 |
| 5 | Rekayasa Aplikasi Peramalan Produk Kosmetik Menggunakan Metode Adaptive Response Rate Single Exponential Smoothing | 39.26 |

Pada salah satu pengujian, sistem menghasilkan nilai similarity tertinggi sebesar 79.36%. Hasil tersebut menunjukkan bahwa sistem mampu mengidentifikasi judul-judul yang memiliki keterkaitan topik dengan judul yang sedang dianalisis.

### C. Evaluasi Pengaruh Preprocessing

Evaluasi dilakukan untuk mengetahui pengaruh setiap tahapan preprocessing terhadap hasil similarity yang dihasilkan oleh metode TF-IDF dan Cosine Similarity. Pengujian dilakukan menggunakan 50 judul uji yang terdiri dari 25 judul skripsi periode sebelumnya dan 25 judul pengajuan terbaru.

Setiap judul diuji menggunakan empat skenario preprocessing, yaitu raw text, cleaning, stopword removal, dan stemming. Nilai similarity tertinggi dari setiap skenario kemudian dihitung rata-ratanya untuk memperoleh gambaran pengaruh preprocessing terhadap keseluruhan data uji.

**Tabel 2.** Hasil Evaluasi Pada 25 Judul Skripsi Periode Sebelumnya.

| Tahap | Rata-rata Similarity (%) |
|---|---|
| Raw Text | 36.89 |
| Cleaning | 37.07 |
| Stopword Removal | 36.61 |
| Stemming | 38.28 |

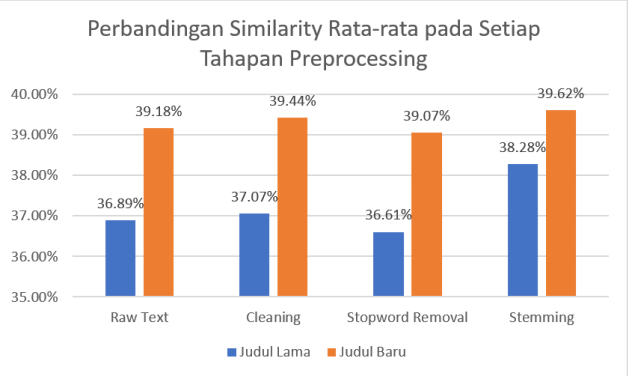
**Tabel 3.** Hasil Evaluasi Pada 25 Judul Pengajuan Terbaru.

| Tahap | Rata-rata Similarity (%) |
|---|---|
| Raw Text | 39.18 |
| Cleaning | 39.44 |
| Stopword Removal | 39.07 |
| Stemming | 39.62 |

**Tabel 4.** Perbandingan Similarity Hasil Evaluasi

| Tahap 4 | Judul Lama (%) | Judul Baru (%) |
|---|---|---|
| Raw Text | 36.89% | 39.18% |
| Cleaning | 37.07% | 39.44% |
| Stopword Removal | 36.61% | 39.07% |
| Stemming | 38.28% | 39.62% |

Gambar 4 menunjukkan perbandingan nilai similarity rata-rata pada setiap tahapan preprocessing untuk kelompok judul lama dan judul baru. Terlihat bahwa tahapan stemming menghasilkan nilai similarity tertinggi pada kedua kelompok data, sedangkan stopword removal menghasilkan nilai similarity yang relatif lebih rendah dibandingkan tahapan lainnya.




Gambar 4. Perbandingan Similarity Rata-rata pada Setiap Tahapan Preprocessing

### D. Pembahasan Hasil Evaluasi

Berdasarkan hasil evaluasi pada dua kelompok data uji, terlihat bahwa tahapan preprocessing memberikan pengaruh terhadap nilai similarity yang dihasilkan. Pada kedua kelompok data, tahapan cleaning menghasilkan peningkatan similarity dibandingkan raw text. Hal ini menunjukkan bahwa proses normalisasi huruf dan pembersihan karakter membantu menghasilkan representasi dokumen yang lebih konsisten.

Tahapan stopword removal menghasilkan penurunan similarity yang relatif kecil pada kedua kelompok data. Pada data judul lama, similarity menurun dari 37.07% menjadi 36.61%, sedangkan pada data judul baru menurun dari 39.44% menjadi 39.07%. Hasil tersebut menunjukkan bahwa daftar stopword yang digunakan tidak menghilangkan istilah utama bidang teknologi informasi sehingga pengaruhnya terhadap similarity relatif kecil.

Tahapan stemming menghasilkan nilai similarity rata-rata tertinggi pada kedua kelompok data. Pada data judul lama, similarity meningkat menjadi 38.28%, sedangkan pada data judul baru mencapai 39.62%. Hasil ini menunjukkan bahwa proses pengubahan kata ke bentuk dasar mampu meningkatkan kesamaan term antar dokumen sehingga menghasilkan nilai similarity yang lebih tinggi.

Selain itu, rata-rata similarity pada kelompok judul baru lebih tinggi dibandingkan kelompok judul lama pada seluruh tahapan preprocessing. Kondisi ini menunjukkan bahwa judul pengajuan terbaru memiliki keterkaitan yang lebih besar dengan koleksi judul skripsi yang terdapat pada database. Temuan tersebut mengindikasikan bahwa topik penelitian yang diajukan mahasiswa menunjukkan kecenderungan memiliki kedekatan dengan penelitian yang telah tersimpan pada database skripsi.

Berdasarkan hasil evaluasi tersebut dapat disimpulkan bahwa preprocessing memberikan pengaruh terhadap proses pengukuran similarity dokumen. Di antara seluruh tahapan preprocessing yang diuji, stemming menghasilkan nilai similarity rata-rata tertinggi pada dataset penelitian. Meskipun

peningkatannya relatif kecil dibandingkan tahapan lainnya, proses stemming mampu menyatukan berbagai variasi bentuk kata ke dalam satu bentuk dasar sehingga jumlah term yang memiliki kesamaan antarjudul menjadi lebih konsisten. Sebagai contoh, kata menggunakan, penggunaan, dan digunakan direpresentasikan menjadi bentuk dasar guna. Penyatuan variasi kata tersebut memungkinkan sistem mengenali hubungan antardokumen dengan lebih baik sehingga nilai similarity cenderung meningkat.

## IV. KESIMPULAN

Penelitian ini berhasil menganalisis pengaruh preprocessing terhadap similarity judul skripsi menggunakan metode TF-IDF dan Cosine Similarity. Pengujian dilakukan terhadap 480 data judul skripsi bidang teknologi informasi sebagai data pembanding dan 50 judul uji yang digunakan untuk mengevaluasi setiap tahapan preprocessing. Selain menghasilkan sistem berbasis web untuk deteksi kemiripan judul skripsi, penelitian ini menunjukkan bahwa preprocessing memberikan perubahan terhadap nilai similarity yang dihasilkan.

Hasil evaluasi terhadap 50 judul uji yang terdiri dari 25 judul skripsi periode sebelumnya dan 25 judul pengajuan terbaru menunjukkan bahwa tahapan preprocessing memberikan pengaruh terhadap nilai similarity yang dihasilkan. Tahap cleaning menghasilkan perubahan nilai similarity yang relatif kecil dibandingkan raw text, sedangkan stopword removal menunjukkan sedikit penurunan nilai similarity pada dataset penelitian. Hasil penelitian menunjukkan bahwa proses stemming menyatukan variasi bentuk kata ke dalam bentuk dasar sehingga representasi term antar dokumen menjadi lebih konsisten. Pada dataset penelitian ini, kondisi tersebut menghasilkan nilai similarity rata-rata tertinggi dibandingkan tahapan preprocessing lainnya.

Berdasarkan seluruh tahapan preprocessing yang diuji, stemming menghasilkan nilai similarity rata-rata tertinggi pada kedua kelompok data uji. Pada kelompok judul lama, similarity meningkat menjadi 38,28%, sedangkan pada kelompok judul baru mencapai 39,62%. Hasil tersebut menunjukkan bahwa proses stemming mampu meningkatkan kesamaan term antar dokumen sehingga menghasilkan pengukuran similarity yang lebih baik.

Selain itu, kelompok judul pengajuan terbaru memiliki nilai similarity rata-rata yang lebih tinggi dibandingkan kelompok judul periode sebelumnya pada seluruh tahapan preprocessing. Temuan ini menunjukkan bahwa judul yang diajukan mahasiswa pada periode terbaru memiliki keterkaitan yang lebih besar dengan koleksi judul skripsi yang telah tersedia pada database.

## REFERENSI

[1] Arsad, A., Hamid, M., dan Santosa, M. “Penerapan Teks Mining Dan *Cosine Similarity* Untuk Menentukan Kesamaan Dokumen Skripsi”. *IJIS, Indonesian Journal on Information System*, vol. 9, no. 1, pp 99-109, 2024.

[2] Azmi, M. “Analisis Tingkat Plagiasi Dokumen Skripsi Dengan Metode Cosine Similarity Dan Pembobotan TF-IDF”. *Jurnal TEKNIMEDIA*. Vol. 2, No. 2, pp 90 – 95, 2021.

[3] Putri, K., Ramadlani,N.A., dan Cahyani, L. “Penerapan Algoritma TF-IDF Dan *Cosine Similarity* Untuk Query Pencarian Soal Mata Pelajaran Sosiologi SMA”. *Jurnal Teknologi Informasi: Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika.* Vol 20 No 1, pp 31 – 46, 2026. doi: 10.47111/ JTI.

[4] Sarimuddin, Azlina n., Anggun dkk. “Analisis Kemiripan Judul Skripsi Menggunakan Pembobotan TF-IDF dan Metode Cosine Similarity Untuk Mencegah Duplikasi”. Jurnal : semanTIK, Vol.11, No.1, pp. 33-42, 2026.

[5] Nico, Budiyanto, U., Fatimah, T. “Implementasi Algoritma Pembobotan TF-IDF dan Cosine Similarity untuk Penetapan Kategori Artikel pada Website Universitas Budi Luhur”. *Jurnal TICOM: Technology of Information and Communication.* Volume 10, Nomor 3, Mei, pp 218-223, 2022.

[6] Widianto, A., Pebriyanto, E., Fitriyanti, Marna. “Document Similarity using Term Frequency-Inverse Document Frequency Representation and Cosine Similarity”. *Journal of Dinda: Data Science, Information Technology, and Data Analytics*. Vol. 4, No. 2, pp 149 – 153, 2024.

[7] Lim, V. I., Fitria, dan Hafid, M. “Implementasi Text Summarization pada Ulasan Aplikasi Mobile JKN Menggunakan TF-IDF dan Cosine Similarity”. *KONVERGENSI, Jurnal Teknologi Informasi dan Komunikasi*, vol. 21 no.1, pp 9–17, 2025. doi: 10.30996/konv.v21i1.12196 .

[8] Tanuwijaya, W., Setiawan, C. E., Irsyad, H. , Rahman, A. “Implementasi TF-IDF dan *Cosine Similarity* untuk Penyaringan Dokumen Berita Program Makan Siang Gratis Pemerintah Indonesia”. *Device : Journal Of Information System, Computer Science And Information Technology,* Vol. 6, No. 2, pp : 322 – 334. 2025.

[9] Sari, H., Ginting, G. L., Zebua, T. “Penerapan Algoritma Text Mining dan TF-IDF Untuk Pengelompokan Topik Skripsi Pada Aplikasi Repository STMIK Budi Darma”. *TIN: Terapan Informatika Nusantara*. Vol 2, No 7, pp 414-432, 2021.

[10] Nasrullah, A., H. “Integrasi Tf-Idf Dan Algoritma *Cosine Similarity* Untuk Deteksi Tingkat Kemiripan Judul Penelitian (Studi Kasus Mahasiswa Fakultas Ilmu Komputer UNISAN Gorontalo)”. *INTEC Journal: Information Technology Education Journal.* Vol. 3, No. 3, pp 113 – 118. 2024.

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

[12] Silalahi, N., Ginting, G. L. “Rekomendasi Berita Berkaitan dengan Menerapkan Algoritma Text Mining

dan TF-IDF". *Bulletin Of Computer Science Research*. Vol 3, No 4, pp 276−282, 2023.

[13] Fahrudin, T. M., Hartanto, M.H., Paramita A.S. "Temu Kembali Informasi Berita Kegiatan Program Studi Studi Menggunakan Algoritma Pembobotan TF-IDF Dan Cosine Similarity". *Prosiding Seminar Nasional Teknologi dan Sistem Informasi (SITASI)*. pp 270 – 279. 2022.

[14] Bakiyev, B. "Method for Determining the Similarity of Text Documents for the Kazakh language, Taking Into Account Synonyms: Extension to TF-IDF ". *Smart Information Systems and Technologies (SIST).* 2022. doi : 10.1109/SIST54437.2022.9945747.