BETRIK
BESEMAH TEKNOLOGI INFORMASI & KOMPUTER
Editor Office : Pusat Penelitian & Pengabdian Pada Masyarakat (PPPM) ITPA
Phone : 0857-9716-9578
email : betrikitpa@itpa.ac.id
ISSN : 2339 - 1871

Pendekatan Model TF-IDF dan *Cosine Similarity* Pada Rekomendasi Dosen Pembimbing Skripsi

Heki Aprianto
Program Studi Administrasi Kesehatan, Sekolah Tinggi Ilmu Kesehatan Budi Mulia Sriwijaya, Indonesia
Sur-el :apriantoheki@gmail.com

Penulis Korespondensi: Heki Aprianto, apriantoheki@gmail.com

Abstrak : Rekomendasi pemilihan dosen pembimbing skripsi ditentukan berdasarkan kesesuaian topik penelitian mahasiswa dengan kompetensi keilmuan yang dimiliki dosen. Proses seleksi dosen pembimbing masih mengandalkan pendekatan manual yang didasarkan pada preferensi subjektif atau ketersediaan dosen sehingga berisiko menimbulkan ketidaksesuaian antara kompetensi dosen dan topik penelitian. Penelitian ini bertujuan untuk menerapkan pendekatan model *machine learning Term Frequency–Inverse Document Frequency* (TF-IDF) dan *cosine similarity* dalam rekomendasi pemilihan dosen pembimbing guna meningkatkan objektivitas dan efisiensi proses akademik. Metode yang digunakan meliputi pengumpulan data berupa 25 dosen dengan masing-masing minimal 5 publikasi sehingga diperoleh total 125 dokumen berupa judul dan abstrak penelitian, serta 40 data topik skripsi mahasiswa sebagai data uji pada STIKes Budi Mulia Sriwijaya. Tahapan pengolahan data mencakup pra-pemrosesan teks berupa *case folding*, *tokenization, stopword removal,* dan *stemming.* Hasil analisis menunjukkan bahwa nilai bobot TF-IDF tertinggi sebesar 95,81 yang mengindikasikan adanya kata kunci yang sangat dominan dan spesifik. Sementara itu, nilai cosine similarity tertinggi sebesar 0,77 menunjukkan tingkat kemiripan yang kuat antara topik penelitian mahasiswa dan bidang keilmuan dosen. Visualisasi hasil dalam bentuk heatmap memperlihatkan adanya pengelompokan dokumen berdasarkan tingkat kemiripan topik.

Kata kunci : *cosine similarity,* kesamaan, pemrosesan teks, rekomendasi dosen, TF-IDF

Abstract *: The recommendation for selecting thesis supervisors is determined based on the alignment between students' research topics and the academic expertise of lecturers. However, the current supervisor selection process still relies on manual approaches based on subjective preferences or lecturer availability, which may lead to mismatches between lecturer competencies and research topics. This study aims to implement a machine learning approach using the Term Frequency–Inverse Document Frequency (TF-IDF) and cosine similarity methods to improve the objectivity and efficiency of the supervisor recommendation process. The method used involves data collection from 25 lecturers, each having at least five publications, resulting in a total of 125 documents in the form of research titles and abstracts. Additionally, 40 student thesis topics were used as testing data at STIKes Budi Mulia Sriwijaya. The data processing stages include text preprocessing, such as case folding, tokenization, stopword removal, and stemming. The results show that the highest TF-IDF weight reaches 95.81, indicating the presence of highly dominant and specific keywords. Meanwhile, the highest cosine similarity value of 0.77 indicates a strong level of similarity between students' research topics and lecturers' areas of expertise. The results are further visualized using a heatmap, which reveals clusters of documents based on topic similarity.*

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

Keywords: *cosine similarity, text processing, similarity, supervisor recommendation, TF-IDF,*

1. PENDAHULUAN

Pemilihan dosen pembimbing skripsi merupakan salah satu tahapan penting dalam proses akademik yang berpengaruh terhadap kualitas dan keberhasilan penelitian mahasiswa. Idealnya, pemilihan pembimbing dilakukan berdasarkan kesesuaian antara topik penelitian mahasiswa dengan kompetensi keilmuan dosen. Pemilihan yang tepat tidak hanya memastikan mahasiswa mendapatkan bimbingan substantif yang berkualitas, tetapi juga menciptakan hubungan mentoring yang efektif dan mendukung [1]. Namun, pada praktiknya proses ini masih sering dilakukan secara manual berdasarkan preferensi subjektif maupun ketersediaan dosen sehingga berpotensi menimbulkan ketidaksesuaian bidang keahlian dengan topik penelitian. Dampak dari permasalahan ini diantaranya terhambatnya proses penyelesaian skripsi, kesulitan komunikasi akademik serta rendahnya kualitas hasil penelitian.

Riset menunjukkan bahwa ketidakcocokan bidang keahlian merupakan salah satu faktor utama penyebab keterlambatan penyelesaian skripsi dan meningkatnya tingkat stres akademik pada mahasiswa [2] sehingga banyak institusi mulai mengadopsi sistem penunjukan atau pencarian pembimbing berbasis data diantaranya menganalisis kesesuaian topik penelitian dengan kompetensi dosen dalam konteks penentuan reviewer profil publikasi dosen dan pemetaan keahlian menggunakan teknik text mining secara lebih objektif dibandingkan pendekatan manual sehingga memiliki kontribusi praktis dalam mendukung pengambilan keputusan berbasis data [3]. Oleh karena itu, dibutuhkan sistem yang mampu memberikan rekomendasi dosen pembimbing skripsi berdasarkan kecocokan bidang penelitian mahasiswa dengan keahlian dosen yang dimiliki.

Beberapa penelitian terdahulu telah mengkaji penggunaan metode TF-IDF dan cosine similarity dalam pengolahan teks. Penelitian yang dilakukan oleh [4] mengembangkan sistem rekomendasi dosen pembimbing dengan metode cosine similarity yang mampu mengukur tingkat kemiripan topik penelitian, namun belum memanfaatkan pembobotan kata seperti TF-IDF sehingga representasi teks masih terbatas. Selanjutnya, penelitian oleh [5] menggunakan cosine similarity dalam pemilihan dosen pembimbing berdasarkan judul skripsi mahasiswa, namun masih terbatas pada penggunaan data judul tanpa mempertimbangkan abstrak penelitian dosen. Selain itu, pada penelitian [6] mengembangkan sistem rekomendasi berbasis cosine similarity, tetapi belum mengintegrasikan teknik pembobotan kata maupun analisis data yang lebih komprehensif.

Berdasarkan penelitian-penelitian tersebut, terdapat beberapa kesenjangan (*research gap*) diantaranya sebagian besar penelitian hanya menggunakan cosine similarity tanpa dikombinasikan dengan metode pembobotan kata seperti TF-IDF, keterbatasan representasi data yang umumnya hanya menggunakan judul tanpa memanfaatkan abstrak penelitian secara menyeluruh serta belum adanya integrasi yang optimal antara data topik penelitian mahasiswa dan kompetensi dosen dalam satu sistem rekomendasi yang komprehensif.

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

All author: Heki Aprianto

Penelitian ini mengusulkan pendekatan sistem rekomendasi dosen pembimbing skripsi dengan mengintegrasikan pendekatan *content-based filtering* dengan model TF-IDF (Term Frequency – Inverse Document Frequency) dan cosine similarity menggunakan data yang lebih representatif, yaitu 25 dosen dengan total 125 dokumen berupa judul dan abstrak penelitian serta 40 data topik skripsi mahasiswa. Selain itu, penelitian ini dilengkapi dengan visualisasi dalam bentuk heatmap untuk memperjelas pola kemiripan antar dokumen. Keterbaruan (novelty) penelitian ini terletak pada integrasi metode pembobotan kata dan pengukuran kemiripan dalam konteks rekomendasi dosen pembimbing skripsi, penggunaan data judul dan abstrak secara bersamaan serta penyajian visualisasi untuk meningkatkan interpretasi hasil rekomendasi.

Pendekatan TF-IDF (*Term Frequency – Inverse Document Frequency*) berfungsi untuk mengekstraksi bobot suatu kata dari sekumpulan dokumen sehingga kata yang sering muncul dalam dokumen tertentu, tetapi jarang muncul pada keseluruhan koleksi dokumen, memiliki bobot yang lebih tinggi melalui representasi dokumen penelitian mahasiswa maupun profil publikasi dosen dalam bentuk vektor numerik yang bermakna [7]. Setelah dokumen direpresentasikan dalam bentuk vektor, proses selanjutnya dilakukan pengukuran tingkat kesamaan antara dua dokumen dengan Pendekatan *Cosine Similarity,* yang mengukur kedekatan sudut antar dua vektor dalam ruang multidimensi, semakin kecil sudut yang terbentuk, semakin tinggi tingkat kesamaan antara dokumen mahasiswa dengan dokumen dosen.

Sekolah Tinggi Ilmu Kesehatan (STIKes) Budi Mulia Sriwijaya merupakan institusi pendidikan tinggi di bidang kesehatan yang berlokasi di Palembang, Sumatera Selatan, Indonesia. STIKes Budi Mulia Sriwijaya menawarkan berbagai program pendidikan di bidang kesehatan yang dirancang untuk membekali mahasiswa dengan pengetahuan teoritis dan keterampilan praktis melalui kegiatan akademik dan praktik lapangan. Pada konteks penelitian ini, permasalahan yang muncul terkait pemilihan dosen pembimbing skripsi pada STIKes Budi Mulia Sriwijaya bukan hanya permasalahan administratif akademik, melainkan juga merupakan fenomena yang dapat dikaji dengan perspektif ilmu komputer, khususnya pada bidang information retrieval, machine learning dan sistem rekomendasi. Konsep dasar ilmu pengetahuan menekankan bahwa setiap fenomena perlu didekati melalui proses observasi, pengukuran dan pemodelan agar dapat diperoleh pemahaman yang rasional serta solusi yang efektif [8].

Fenomena yang terjadi dalam bidang akademik pada STIKes Budi Mulia Sriwijaya adalah kesulitan dalam menentukan kesesuaian antara dosen pembimbing skripsi dengan bidang penelitian mahasiswa. Hal ini dapat menimbulkan masalah dalam proses bimbingan, baik dari sisi keilmuan, kesesuaian bidang riset, maupun keterbatasan jumlah dosen yang dapat menangani mahasiswa dalam satu periode. Permasalahan ini menjadi semakin kompleks ketika tidak ada sistem yang secara objektif menghubungkan topik penelitian mahasiswa dengan bidang keahlian dosen. Dari sudut pandang sistem informasi, fenomena ini merupakan masalah pencocokan informasi (*matching problem*) yang membutuhkan pendekatan komputasional agar dapat diselesaikan secara optimal.

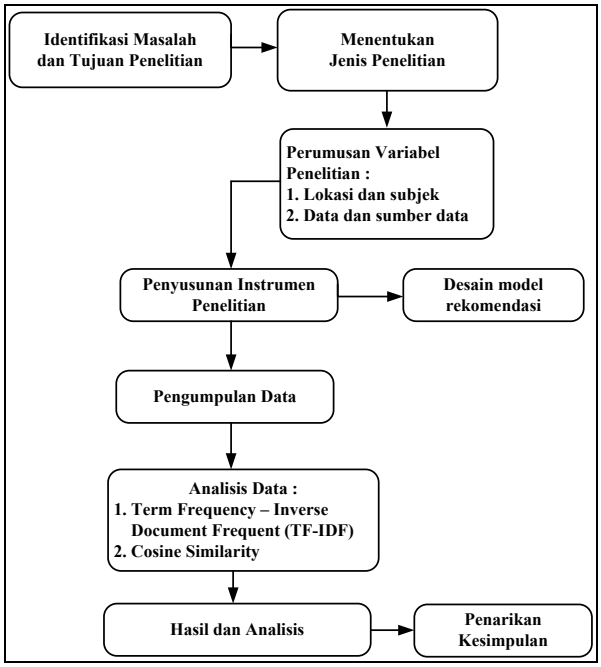
Manfaat dari penelitian ini diharapkan tidak hanya mendukung bidang akademik dalam menentukan dosen pembimbing skripsi yang sesuai, tetapi juga membantu STIKes Budi Mulia Sriwijaya dalam mendistribusikan beban bimbingan secara lebih merata. Selain itu, hasil penelitian ini dapat menjadi

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

referensi dalam pengembangan sistem rekomendasi serupa di bidang lain, seperti rekomendasi mata kuliah, pemilihan topik penelitian maupun kolaborasi akademik.

2. METODOLOGI PENELITIAN

Metode penelitian yang dilakukan dalam penelitian untuk merancang sistem rekomendasi dosen pembimbing skripsi dengan menerapan model TF-IDF dan *cosine similarity* digambarkan dalam diagram alur penelitian gambar 1 dibawah ini.



Gambar 1. Diagram Alur Penelitian

Penelitian ini menggunakan beberapa tahapan yang digambarkan dalam Gambar 1 melalui pendekatan kuantitatif dengan metode eksperimen. Pendekatan kuantitatif dipilih karena penelitian ini berfokus pada pengolahan data berbasis teks dengan model matematis (TF-IDF) dan pengukuran kemiripan antar dokumen (cosine similarity). Sedangkan metode eksperimen digunakan karena penelitian melibatkan proses implementasi sistem rekomendasi serta pengujian efektivitasnya melalui pengukuran tingkat akurasi dan relevansi hasil rekomendasi.

Tahap selanjutnya adalah menentukan lokasi penelitian yang dilakukan di Perguruan Tinggi STIKes Budi Mulia Sriwijaya dengan melibatkan dua jenis data utama, yaitu:

1. Data mahasiswa berupa judul atau topik penelitian skripsi yang diajukan sebanyak 40 data mahasiswa untuk digunakan sebagai data uji dalam sistem rekomendasi.
2. Data dosen pembimbing berupa nama dosen, serta dokumen penelitian yang terdiri dari judul dan abstrak publikasi ilmiah. Data yang digunakan mencakup 25 dosen dengan masing-masing minimal 5 publikasi sehingga diperoleh total 125 dokumen penelitian. Subjek penelitian adalah mahasiswa tingkat akhir yang sedang menyusun skripsi serta dosen yang memiliki kewenangan sebagai pembimbing.

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

Pada penelitian ini jenis data yang digunakan terdiri dari : data primer berupa kuesioner mahasiswa yang ditampilkan pada Tabel 1 tentang kesulitan memilih dosen pembimbing dan wawancara dengan dosen terkait distribusi bimbingan dan kesesuaian topik penelitian.

Tabel 1. Instrumen Kuesioner Mahasiswa

No	Variabel		Indikator	Pernyataan	Skala
1	Kesulitan Pemilihan Pembimbing	1	Tingkat kesulitan	Saya mengalami kesulitan dalam menentukan dosen pembimbing skripsi	Likert 1–5
		2	Ketersediaan informasi	Informasi mengenai bidang keahlian dosen sulit diperoleh	Likert 1–5
2	Kesesuaian Topik	3	Kesesuaian topik dengan dosen	Saya kesulitan mencocokkan topik penelitian dengan keahlian dosen	Likert 1–5
		4	Pemahaman bidang dosen	Saya memahami bidang keilmuan dosen pembimbing yang tersedia	Likert 1–5
3	Proses Pemilihan	5	Metode pemilihan	Pemilihan dosen pembimbing masih berdasarkan rekomendasi teman	Likert 1–5
		6	Subjektivitas	Pemilihan dosen pembimbing cenderung subjektif	Likert 1–5
4	Kebutuhan Sistem	7	Kebutuhan teknologi	Saya membutuhkan sistem yang dapat membantu merekomendasikan dosen pembimbing	Likert 1–5
		8	Efektivitas sistem	Sistem rekomendasi dapat membantu memilih dosen yang sesuai dengan topik penelitian	Likert 1–5
5	Efisiensi Proses	9	Waktu pemilihan	Proses pemilihan dosen pembimbing saat ini memerlukan waktu yang lama	Likert 1–5
		10	Kemudahan proses	Saya menginginkan proses pemilihan dosen pembimbing yang lebih mudah dan cepat	Likert 1–5

Skala pengukuran menggunakan skala Likert dengan rentang nilai 1–5, yaitu:

1 = Sangat Tidak Setuju, 2 = Tidak Setuju, 3 = Netral, 4 = Setuju, 5 = Sangat Setuju.

Selanjutnya, data sekunder berupa dokumen akademik atau data penelitian seperti daftar dosen, bidang keilmuan, judul dan abstrak pada publikasi ilmiah disajikan pada Tabel 2 yang berisi sebagian sampel data sekunder tersebut untuk memudahkan penyajian dan keterbacaan data. Secara keseluruhan, data sekunder terdiri dari 25 dosen dengan total 125 dokumen data penelitian.

Tabel 2. Data sekunder berupa dokumen akademik

No	Nama Dosen	Bidang Keilmuan	Judul Jurnal	Abstrak
1	Venny Mayumi Gultom	Kesehatan Masyarakat	Evaluasi Tingkat Kepuasan Pasien Terhadap Pelayanan Dokter	Tujuan penelitian ini untuk mengetahui tingkat kepuasan layanan dokter di fasilitas kesehatan dalam sistem pelayanan asuransi kesehatan BPJS bidang kesehatan. Pendekatan yang digunakan dalam penelitian ini adalah pendekatan kuantitatif..
			Identifikasi Penyakit Tidak Menular (PTM) Melalui Kadar Gula Darah,	Tujuan penelitian ini untuk mengidentifikasi PTM melalui kadar gula darah, asam urat dan kolesterol masyarakat. Penelitian

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

No	Nama Dosen	Bidang Keilmuan	Judul Jurnal	Abstrak
			Asam Urat dan Kolesterol	ini diambil menggunakan teknik purposive sampling ditentukan yaitu sejumlah 83 orang dengan kriteria inklusi yang ditentukan.
2	Siti Hajrianti	Terapan Kebidanan	Deteksi Anemia Pada Ibu Hamil Mengunakan Metode Non Invasif Berbasis Kecerdasan	Penelitian ini bertujuan untuk membuat alat deteksi anemia pada ibu hamil non invasif berbasis kecerdasan artifisial. Hasil penelitian menunjukan bahwa alat nilai AUC sebesar 91% dan nilai sensitivitas alat sebesar 94%
3	Tirta Anggraini	Kesehatan Terapan	Faktok-Faktor yang berhubungan Dengan Pemberian ASI Eksklusif Pada Bayi 6-12 bulan di Puskesmas Sosial Palembang Tahun 2019	Penelitian ini bertujuan untuk mengetahui Faktor-Faktor apa saja yang berhubungan dengan Pemberian ASI Eksklusif di puskesmas social Palembang Tahun 2019. Metode penelitian menggunakan desain cross sectional dan menggunakan teknik wawancara dengan menggunakan kuesioner,
			Pengetahuan Ibu Hamil tentang Kontrasepsi Metode Amenore Laktasi (MAL)	penelitian ini adalah untuk mengetahui Deskripsi Pengetahuan Ibu Hamil tentang Kontrasepsi Amenore Laktasi (LAM) di Klinik Medika Budi Mulia, Palembang pada tahun 2025. Desain penelitian deskriptif. Populasi dalam penelitian ini adalah ibu hamil yang berkunjung ke Klinik Budi Mulia.
4	Ari Darmansyah	Administrasi Kesehatan	Gambaran Pemberian Hidroterapi Air Hangat dan Eco Enzyme Terhadap Nilai Tekanan Darah	Penelitian ini bertujuan untuk mengidentifikasi perbedaan tekanan darah sebelum dan setelah pemberian terapi rendam kaki menggunakan air hangat dan eco enzyme, dengan pengambilan sampel dilakukan menggunakan teknik purposive sampling dengan kriteria inklusi berusia 20-60 tahun,
			Hubungan Mutu Pelayanan Kesehatan dan Konerja Tenaga Kesehatan dengan Tingkat Kepuasan pasien BPJS Rawat Jalan.	Tujuan dari penelitian ini untuk mengetahui hubungan antara kualitas pelayanan di departemen pendaftaran dengan kepuasan rawat jalan di X Palembang Health Care. sampel yang tidak disengaja dengan populasi 107.520 pasien dan sampel diambil dari 100 responden menggunakan perhitungan rumus Slovin.

Teknik pengumpulan data dilakukan dengan beberapa metode diantaranya : studi pustaka dengan mengkaji literatur terkait model TF-IDF dan cosine similarity pada rekomendasi pemilihan dosen pembimbing skripsi [9], [10] yang mengidentifikasi alur proses pemilihan dosen pembimbing skripsi pada

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

program studi Administrasi Kesehatan, mengumpulkan data publikasi dosen dan topik penelitian mahasiswa, kuesioner dari mahasiswa dan dosen. Data yang telah terkumpul kemudian diproses dalam tahap analisis data, yang dalam diagram ini menggunakan dua teknik komputasional khusus untuk data tekstual atau yang dapat direpresentasikan sebagai teks. teknik pertama adalah *Term Frequency – Inverse Document Frequency* (TF-IDF), sebuah metode statistik untuk mengevaluasi pentingnya suatu kata dalam sebuah dokumen relatif terhadap sekumpulan dokumen. TF-IDF menghasilkan pembobotan kata-kata kunci, yang mengubah data teks mentah menjadi representasi numerik (vektor) yang siap diolah [11]. Teknik kedua adalah cosine similarity, yang digunakan untuk mengukur tingkat kemiripan antara dua vektor dokumen tersebut. Dalam konteks penelitian ini, cosine similarity kemungkinan besar berfungsi untuk membandingkan dokumen satu sama lain, mengelompokkannya, atau mencocokkannya dengan query tertentu sebagai bagian dari mekanisme model rekomendasi yang dirancang sebelumnya [12].

Tahap akhir penelitian berfokus pada sintesis dan penarikan makna dari temuan analisis. Hasil dari perhitungan TF-IDF dan cosine similarity diinterpretasikan secara mendalam pada bagian hasil dan analisis. Di sini, peneliti tidak hanya melaporkan angka atau ukuran kemiripan, tetapi juga menjelaskan pola, hubungan dan implikasi dari temuan tersebut terkait dengan masalah penelitian awal. Pada bagian kesimpulan, peneliti menjawab pertanyaan penelitian dan merangkum temuan utama secara ringkas dan menyoroti pencapaian tujuan penelitian. Kesimpulan ini sekaligus mengikat kembali identifikasi masalah hingga hasil analisis dan diikuti dengan saran untuk penelitian lanjutan

3. HASIL DAN PEMBAHASAN

3.1 Hasil

Hasil yang diuraikan dalam penelitian ini berupa hasil perhitungan metode content based filtering untuk merekomendasikan [13] nama dosen pembimbing skripsi pada STIKes Budi Mulia Sriwijaya.

3.1.1 Tahap pengumpulan data dan pembuatan *dataset*

Data merupakan komponen penting dalam rekomendasi pemilihan dosen pembimbing skripsi pada penelitian ini. Pengumpulan data dilakukan secara sistematis untuk memastikan kualitas dan relevansi data yang digunakan. Dataset penelitian ini berisi kumpulan data yang disusun secara terstruktur dalam format tabel, di mana setiap barisnya mewakili satu entitas seperti data dosen beserta publikasinya dan setiap kolomnya menggambarkan atribut atau karakteristik dari entitas tersebut, seperti nama, judul publikasi dan abstrak yang ditampilkan pada Tabel 2.

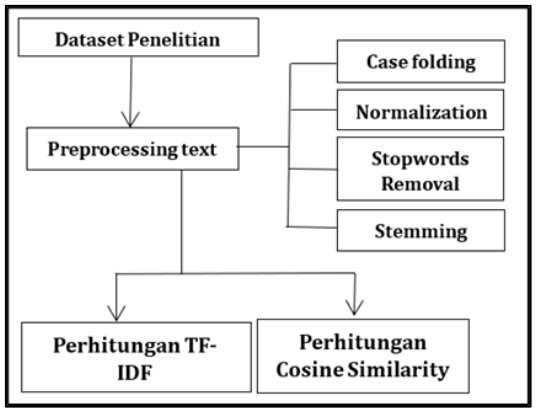
Dalam konteks penelitian ini, dataset berfungsi memuat informasi tentang dosen dan karya ilmiah nya, yang dikumpulkan dari sumber seperti google scholar dan repository institusi. Data tersebut kemudian melalui proses pembersihan dan validasi untuk memastikan konsistensi dan keakuratannya sebelum disimpan dalam format csv di excel, yang memudahkan proses analisis dan integrasi dengan sistem perhitungan yang di lakukan di *google colab*.

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

All author: Heki Aprianto

3.1.2 Tahap Teks *Preprocessing*

Pada penelitian ini, dilakukan tahap teks *preprocessing* untuk membersihkan dan menormalkan data sebelum dilakukan perhitungan TF-IDF. Tahapan ini bertujuan untuk mengurangi noise serta meningkatkan kualitas representasi teks. Adapun tahapan preprocessing yang dilakukan ditampilkan pada Gambar 2 yang berisi tahap preprocessing text pada penelitian ini.



Gambar 2. Alur Tahap Preprocessing Penelitian

1. *Case folding*

   Tahap ini dilakukan untuk mengubah seluruh huruf dalam teks menjadi huruf kecil (lowercase), sehingga tidak terjadi perbedaan antara huruf kapital dan huruf kecil.

2. *Normalization*

   Pada tahap ini dilakukan proses normalisasi kata, seperti penyesuaian kata tidak baku atau singkatan menjadi bentuk baku agar konsisten dalam analisis.

3. *Stopwords Removal*

   Tahap ini bertujuan untuk menghilangkan kata-kata umum yang tidak memiliki makna penting (*stopwords*), seperti “dan”, “di”, “yang”, sehingga hanya menyisakan kata-kata yang relevan.

4. *Stemming*

   Tahap stemming dilakukan untuk mengubah kata menjadi bentuk dasarnya dengan menghilangkan imbuhan seperti awalan dan akhiran.

Untuk memperjelas proses preprocessing yang dilakukan, berikut disajikan sampel data transformasi teks dari salah satu dokumen penelitian pada Tabel 3 :

Tabel 3. Sampel data transformasi teks dari salah satu dokumen penelitian

Tahapan	Hasil Teks
Teks Asli	Penelitian Ini Bertujuan Untuk Menganalisis Pengaruh Terapi Rendam Kaki Menggunakan Air Hangat Dan Eco Enzyme Terhadap Tekanan Darah Pada Penderita Hipertensi
Case Folding	penelitian ini bertujuan untuk menganalisis pengaruh terapi rendam kaki menggunakan air hangat dan eco enzyme terhadap tekanan darah pada penderita hipertensi
Normalization	penelitian ini bertujuan untuk menganalisis pengaruh terapi rendam kaki menggunakan air hangat dan eco enzyme terhadap tekanan darah pada penderita hipertensi

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

Tahapan	Hasil Teks
Stopword Removal	penelitian bertujuan menganalisis pengaruh terapi rendam kaki menggunakan air hangat eco enzyme tekanan darah penderita hipertensi
Stemming	teliti tuju analisis pengaruh terapi rendam kaki guna air hangat eco enzyme tekan darah derita hipertensi

Berdasarkan Tabel 3, proses *preprocessing* secara bertahap mengubah teks dari bentuk awal menjadi lebih terstruktur dan ringkas. Proses ini menghilangkan kata-kata yang tidak relevan serta menyederhanakan bentuk kata sehingga menghasilkan representasi teks yang lebih optimal untuk proses pembobotan menggunakan TF-IDF dan perhitungan *cosine similarity*.

3.2 Pembahasan

3.2.1 TF-IDF (*Term Frequency–Inverse Document Frequency*)

Tahap awal dalam penelitian ini adalah melakukan pembobotan kata menggunakan metode TF-IDF untuk merepresentasikan dokumen dalam bentuk numerik. Metode ini digunakan untuk menentukan tingkat kepentingan suatu kata dalam dokumen berdasarkan frekuensi kemunculan dan tingkat keunikannya terhadap seluruh dokumen.

Perhitungan TF-IDF terdiri dari dua komponen utama, yaitu Term Frequency (TF) dan Inverse Document Frequency (IDF). TF digunakan untuk mengukur seberapa sering suatu kata muncul dalam dokumen, sedangkan IDF digunakan untuk mengukur seberapa unik kata tersebut dalam keseluruhan dokumen.

$$\mathrm{TF}(t, d) = \frac{\mathrm{ft,d}}{\sum t' \in d \int t', d} \quad \text{..... (1)}$$

$$\mathrm{IDF}(t) = \log\left(\frac{N}{\mathrm{df}(t)}\right) \quad \text{..... (2)}$$

$$\text{TF-IDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t) \quad \text{... (3)}$$

Keterangan :

$f_{t,d}$ adalah jumlah kemunculan term dalam dokumen

N adalah jumlah total dokumen

$\mathrm{df}(t)$ adalah jumlah dokumen yang mengandung term tersebut

3.2.1.1 Perhitungan TF-IDF (*Term Frequency-Inverse Document Frequency*)

Proses perhitungan TF-IDF pada penelitian ini dilakukan melalui beberapa tahap, yaitu:

1. Melakukan preprocessing teks yang meliputi case folding, tokenization, stopword removal dan stemming.
2. Menghitung nilai TF untuk setiap kata pada masing-masing dokumen.
3. Menghitung nilai IDF berdasarkan distribusi kata pada seluruh dokumen.
4. Mengalikan nilai TF dan IDF untuk memperoleh bobot TF-IDF setiap kata.
5. Menjumlahkan seluruh bobot TF-IDF dalam satu dokumen untuk memperoleh total bobot TF-IDF dokumen.

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

3.2.1.2 Penerapan Perhitungan TF-IDF

Penerapan perhitungan TF-IDF dilakukan pada dokumen abstrak dengan jumlah kata sebanyak 173. Setiap kata dalam dokumen dihitung nilai TF berdasarkan frekuensi kemunculannya, kemudian dihitung nilai IDF berdasarkan distribusi kata tersebut pada seluruh dokumen. Nilai TF dan IDF kemudian dikalikan untuk memperoleh bobot TF-IDF masing-masing kata.

Sebagai ilustrasi, jika suatu kata muncul sebanyak 8 kali dalam dokumen dengan total 173 kata, dan kata tersebut muncul pada 20 dari total 125 dokumen, maka diperoleh nilai TF sebesar 8/173 = 0,0462 dan nilai IDF sebesar log(125/20) ≈ 0,796. Dengan demikian, nilai TF-IDF untuk kata tersebut adalah sebesar 0,0367.

Nilai TF-IDF yang diperoleh merupakan bobot untuk setiap kata. Selanjutnya, seluruh bobot TF-IDF dalam dokumen dijumlahkan untuk menghasilkan total bobot TF-IDF dokumen, sebagaimana dirumuskan sebagai berikut : $\text{Total TF} - \text{IDF Dokumen} = \sum \text{TF} - \text{IDF}$.

3.2.1.3 Hasil perhitungan TF-IDF

Hasil perhitungan TF-IDF pada beberapa dokumen penelitian dosen disajikan pada Tabel 4.

Tabel 4. Hasil Perhitungan TF-IDF

Nama Dosen	Judul Jurnal	Kata Abstrak	Rata-rata TF-IDF	Total Bobot TF-IDF
Venny Mayumi Gultom	Evaluasi Tingkat Kepuasan Pasien Terhadap Pelayanan Dokter Di Puskesmas Sungai Lilin Kabupaten Musi Banyuasin Dalam Rangka Pelaksanaan Sistem Asuransi Kesehatan BPJS	173	0.554	95.81
Ari Darmansyah	Gambaran Pemberian Hidroterapi Air Hangat dan Eco Enzyme Terhadap Nilai Tekanan Darah	251	0.361	90.60
Afwan Syarif	Analisis Faktor Risiko Hipertensi: Studi Kasus Keterkaitan Paritas, Umur dan Indeks Massa Tubuh	227	0.374	84.88
Neni Triana	Hubungan Tingkat Pengetahuan dan Sikap Mahasiswa STIKES Budi Mulia Sriwijaya terhadap Tindakan Pencegahan Demam Berdarah Dangue (DBD)	152	0.539	81.97
Heki Aprianto	Rancang Bangun Website Pengolahan Data Alumni Pada Lembaga Kursus dan Pelatihan PalComTech Sudirman	123	0.663	81.63
Betty Sirait	Pengaruh Pengetahuan dan Dampak Pernikahan Dini di Desa Kerta Dewa Kabupaten Musi Rawas Utara Tahun 2024	247	0.286	70.64
Faradillah	Hubungan Kualitas Pelayanan Administrasi dan Petugas Kesehatan terhadap Kepuasan Pasien	234	0.345	80.65

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

Nilai total bobot TF-IDF diperoleh dari penjumlahan seluruh bobot TF-IDF setiap kata dalam dokumen. Selain itu, ditampilkan nilai rata-rata TF-IDF yang diperoleh dari pembagian total bobot dengan jumlah kata dalam dokumen. Berdasarkan Tabel 4, dokumen dengan nilai total TF-IDF tertinggi sebesar 95,81 menunjukkan bahwa dokumen tersebut memiliki kata-kata kunci yang lebih dominan dan spesifik dibandingkan dokumen lainnya. Sementara itu, nilai rata-rata TF-IDF memberikan gambaran tingkat kepentingan kata secara umum dalam dokumen.

Semakin tinggi nilai TF-IDF, maka semakin besar kontribusi kata-kata dalam merepresentasikan topik penelitian. Hal ini akan berpengaruh pada tahap selanjutnya, yaitu perhitungan cosine similarity, di mana dokumen dengan bobot kata yang kuat akan memiliki peluang lebih besar untuk menghasilkan tingkat kemiripan yang tinggi terhadap topik penelitian mahasiswa.

3.2.2 Perhitungan *Cosine Similarity*

Setelah dilakukan pembobotan dokumen menggunakan TF-IDF, tahap selanjutnya adalah menghitung tingkat kemiripan antara topik penelitian mahasiswa dan dokumen penelitian dosen menggunakan metode cosine similarity. Metode ini mengukur kesamaan dua dokumen berdasarkan sudut antara dua vektor dalam ruang multidimensi.

$$\text{Cosine Similarity}(A, B) = \frac{A.B}{\|A\|\|B\|} \quad \text{..... (4)}$$

Keterangan:

A = vektor TF-IDF dokumen mahasiswa

B = vektor TF-IDF dokumen dosen

Nilai cosine similarity berada pada rentang 0 hingga 1, di mana nilai mendekati 1 menunjukkan tingkat kemiripan yang tinggi, sedangkan nilai mendekati 0 menunjukkan kemiripan yang rendah.

3.2.2.1 Tahapan Perhitungan *Cosine Similarity*

1. Mengubah topik penelitian mahasiswa menjadi vektor TF-IDF.
2. Mengubah dokumen penelitian dosen menjadi vektor TF-IDF.
3. Menghitung nilai cosine similarity antara vektor mahasiswa dan masing-masing dosen.
4. Menghasilkan nilai kemiripan sebagai dasar rekomendasi dosen pembimbing.
5. Penerapan perhitungan cosine similarity

3.2.2.2 Penerapan Perhitungan *Cosine Similarity*

Penerapan perhitungan cosine similarity dilakukan pada salah satu data dosen, yaitu Ari Darmansyah dengan judul penelitian "Gambaran Pemberian Hidroterapi Air Hangat dan Eco Enzyme Terhadap Nilai Tekanan Darah", diperoleh representasi vektor TF-IDF sebagai berikut (hasil ekstraksi dari dokumen) : Vektor mahasiswa (A) = (w_1, w_2, w_3, …, w_n), Vektor dosen (B) = (v_1, v_2, v_3, …, v_n)

Selanjutnya dilakukan perhitungan:

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

1. Dot Product (A · B) : $A \cdot B = \sum(wi \times vi)$
2. Panjang Vektor : $\| A \| = \sqrt{\sum w_i^2}, \| B \| = \sqrt{\sum v_i^2}$
3. Cosine Similarity $= \frac{A.B}{\|A\|\|B\|}$

Berdasarkan hasil perhitungan sistem, diperoleh nilai cosine similarity sebesar 0,63 untuk dosen Ari Darmansyah. Nilai tersebut menunjukkan bahwa terdapat tingkat kemiripan yang cukup antara topik penelitian mahasiswa dengan dokumen penelitian dosen, meskipun tidak setinggi beberapa dosen lainnya.

3.2.2.3 Hasil Perhitungan *Cosine Similarity*

Hasil perhitungan *cosine similarity* antara topik penelitian mahasiswa dan dokumen dosen disajikan pada Tabel 5.

Tabel 5. Hasil perhitungan *cosine similarity*

Nama Dosen	Judul Jurnal	Skor Cosine Similarity
Venny Mayumi Gultom	Evaluasi Tingkat Kepuasan Pasien Terhadap Pelayanan Dokter Di Puskesmas Sungai Lilin Kabupaten Musi Banyuasin Dalam Rangka Pelaksanaan Sistem Asuransi Kesehatan BPJS	0,72
Ari Darmansyah	Gambaran Pemberian Hidroterapi Air Hangat dan Eco Enzyme Terhadap Nilai Tekanan Darah	0,63
Afwan Syarif	Analisis Faktor Risiko Hipertensi: Studi Kasus Keterkaitan Paritas, Umur dan Indeks Massa Tubuh	0,68
Neni Triana	Hubungan Tingkat Pengetahuan dan Sikap Mahasiswa STIKES Budi Mulia Sriwijaya terhadap Tindakan Pencegahan Demam Berdarah Dangue (DBD)	0,75
Heki Aprianto	Rancang Bangun Website Pengolahan Data Alumni Pada Lembaga Kursus dan Pelatihan PalComTech Sudirman	0,51
Betty Sirait	Pengaruh Pengetahuan dan Dampak Pernikahan Dini di Desa Kerta Dewa Kabupaten Musi Rawas Utara Tahun 2024	0,58
Faradillah	Hubungan Kualitas Pelayanan Administrasi dan Petugas Kesehatan terhadap Kepuasan Pasien	0,77

Pada Tabel 5 hanya menampilkan sebagian sampel hasil perhitungan cosine similarity untuk memudahkan penyajian dan keterbacaan data. Secara keseluruhan, perhitungan dilakukan terhadap 25 dosen dengan total 125 dokumen penelitian. Setiap topik mahasiswa dibandingkan dengan seluruh dokumen dosen sehingga hasil rekomendasi tetap mempertimbangkan keseluruhan data.

3.3 Analisis Hasil

Berdasarkan Tabel 5, nilai cosine similarity tertinggi sebesar 0,77 menunjukkan tingkat kemiripan yang paling tinggi antara topik penelitian mahasiswa dan dokumen dosen, sehingga dosen tersebut

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

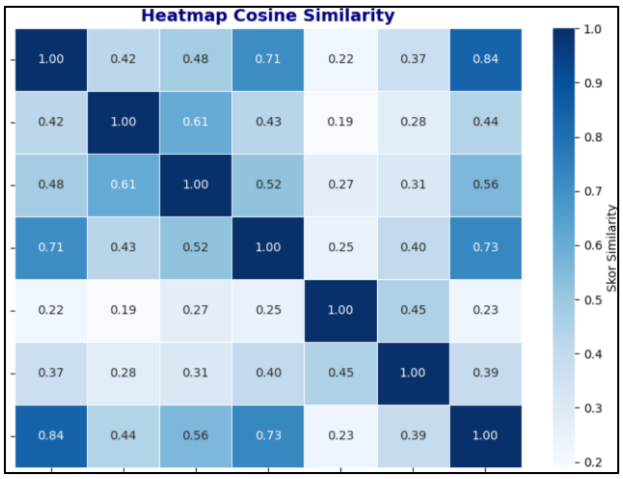
direkomendasikan sebagai pembimbing utama. Nilai lain seperti 0,75 dan 0,72 juga menunjukkan tingkat kesesuaian yang tinggi dan dapat dijadikan alternatif rekomendasi. Sebaliknya, nilai yang lebih rendah seperti 0,51 menunjukkan tingkat kemiripan yang relatif rendah, sehingga kurang direkomendasikan sebagai pembimbing utama.

Keterkaitan antara nilai TF-IDF dan cosine similarity diperoleh dari hasil pembobotan TF-IDF yang telah dilakukan sebelumnya. TF-IDF berperan dalam membentuk vektor numerik dokumen berdasarkan bobot kata, sedangkan cosine similarity digunakan untuk mengukur kedekatan antar vektor tersebut. Semakin tinggi bobot TF-IDF pada kata-kata yang relevan, maka semakin besar kemungkinan nilai cosine similarity yang dihasilkan juga tinggi.

3.4 Visualisasi Heatmap *Cosine Similarity*

Setelah dilakukan perhitungan cosine similarity antara seluruh pasangan dokumen, diperoleh matriks kemiripan (*similarity matrix*) yang berisi nilai kemiripan antar dokumen. Matriks ini merepresentasikan hubungan antar dokumen dalam bentuk numerik, di mana setiap elemen menunjukkan tingkat kemiripan antara satu dokumen dengan dokumen lainnya. Selanjutnya, matriks cosine similarity tersebut divisualisasikan dalam bentuk heatmap untuk memudahkan interpretasi pola kemiripan antar dokumen.

Visualisasi heatmap pada Gambar 3 berfungsi untuk mengidentifikasi kedekatan semantik antar judul dokumen berdasarkan tingkat kemiripannya. Dokumen yang memiliki nilai cosine similarity tinggi akan membentuk pola klaster, sehingga dapat digunakan untuk mengelompokkan topik penelitian yang serupa dan mengarah pada dosen pembimbing dengan keahlian yang paling relevan.



Gambar 3. Visualisasi *heatmap cosine similarity*

Warna pada heatmap merepresentasikan besarnya nilai kemiripan antar dokumen. Warna biru tua menunjukkan nilai similarity yang tinggi (misalnya pada rentang 0,75–0,84), sedangkan warna yang lebih terang menunjukkan nilai kemiripan yang lebih rendah (misalnya 0,20–0,50). Semakin gelap warna yang ditampilkan, maka semakin tinggi tingkat kesamaan topik antar dokumen. Dengan demikian, kombinasi metode TF-IDF dan cosine similarity tidak hanya menghasilkan nilai numerik, tetapi juga dapat

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

divisualisasikan untuk memperlihatkan pola hubungan antar dokumen. Pendekatan ini membentuk dasar model content-based filtering dalam sistem rekomendasi pemilihan dosen pembimbing skripsi.

4. KESIMPULAN

Berdasarkan hasil penelitian mengenai Pendekatan model machine learning TF-IDF dan cosine similarity pada rekomendasi dosen pembimbing skripsi maka dapat disimpulkan bahwa penerapan metode Term Frequency–Inverse Document Frequency (TF-IDF) terbukti efektif dalam mengidentifikasi bobot kata yang paling berpengaruh dalam setiap judul atau abstrak penelitian dosen. Nilai TF-IDF yang tinggi menunjukkan tingkat keunikan dan relevansi istilah terhadap keseluruhan korpus. Selanjutnya, hasil pengukuran cosine similarity menunjukkan kemampuan menentukan tingkat kedekatan semantik antar judul penelitian. Nilai similarity tertinggi diperoleh pada penelitian yang memiliki skor hingga 0,77. Sebaliknya, nilai similarity yang rendah, seperti 0,51. Kemudian hasil visualisasi dalam bentuk heatmap memberikan representasi yang jelas mengenai distribusi tingkat kemiripan antar judul jurnal. Visualisasi ini memudahkan proses analisis pola keterkaitan antar topik penelitian serta memperlihatkan potensi pengelompokan bidang ilmu secara otomatis.

5. UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan dan kontribusi dalam penelitian ini, termasuk rekan dosen dan rekan peneliti, serta responden yang berpartisipasi, sehingga penelitian dan pengembangan dengan Pendekatan Model Machine Learning TF-IDF dan Cosine Similarity Pada Rekomendasi Dosen Pembimbing Skripsi dapat terlaksana dengan baik

DAFTAR RUJUKAN

[1] A. Syafi'i, A. Munir, S. Anam, and S. Suhartono, "Unleashing the power of supervisory feedback in academic writing: Strategies for timely undergraduate thesis completion," *Teflin J.*, vol. 35, no. 2, pp. 330–351, 2024.
[2] S. Chen, L. Zhang, and M. Li, "Doctoral students' self-regulated learning: the roles of academic buoyancy and perceived autonomy support," *Educ. Psychol.*, vol. 45, no. 2, pp. 148–167, 2025.
[3] H. R. Adli, M. Munir, and R. Megasari, "Implementation of Inverse Document Frequency (TF-IDF) and Cosine Similarity Terms in Determining Research Reviewers for Indonesian Education University Lecturers," *J. Comput. Soc.*, vol. 4, no. 2, pp. 73–82.
[4] H. Hairani and M. Mujahid, "Recommendations of Thesis Supervisor using the Cosine Similarity Method," *Sist. J. Sist. Inf.*, vol. 11, no. 3, pp. 646–654, 2022.
[5] R. Andinisari, F. Riski, and K. Auliasari, "Pemilihan Dosen Pembimbing Berdasarkan Judul Skripsi Mahasiswa menggunakan Metode Cosine Similarity," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 7, no. 2, pp. 268–275, 2025.
[6] Z. F. FALAH and S. T. Fajar Suryawan, "Sistem Rekomendasi Pemilihan Dosen Pembimbing Tugas Akhir Dengan Metrik Cosine Similarity." Universitas Muhammadiyah Surakarta, 2021.
[7] M. Rashmi, *Introduction to Information Retrieval Systems*, vol. 3, no. 4. Cambridge university press, 2015. doi: 10.17762/ijritcc2321-8169.150462.
[8] K. Popper, "The Logic of Scientific Discovery," 2012.
[9] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: Techniques, applications, and

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

challenges," *Recomm. Syst. Handb.*, pp. 1–35, 2021.
[10] W. Pedrycz, "The benefits and drawbacks of data mining technologies," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 1, 2020, doi: 10.1002/widm.1344.
[11] N. Azizah and A. F. Rozi, "Sistem Rekomendasi Produk Somethinc Menggunakan Metode Content-based Filtering," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 6, no. 3, pp. 461–468, 2024, doi: 10.47233/jteksis.v6i3.1411.
[12] Dino Akbar Pratondo, "Pengembangan Sistem Rekomendasi Berbasis Content-Based Filtering Pada data Dinamis," *Universitas Islam Negeri Syarif Hidayatullah Jakarta*. Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta, pp. 1–89, 2023. [Online]. Available: https://repository.uinjkt.ac.id/dspace/bitstream/123456789/66813/1/DINO AKBAR PRATONDO-FST.pdf
[13] R. Insan Pratama Siagian, N. Khoiriah, S. Audy Priscilia, M. Raffi Akbar Tanjung, and A. Perdana, "Penerapan Machine Learning Untuk Rekomendasi Film Berdasarkan Preferensi Pengguna," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 4, pp. 5658–5662, 2025, doi: 10.36040/jati.v9i4.13884.

Received: 09-03-2026 | Accepted: 25-04-2026 | Published Online: 30-04-2026

All author: Heki Aprianto