Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi
https://ojs.stmik-banjarbaru.ac.id/index.php/jutisi/index
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0893

Aplikasi Rekomendasi Dosen Pembimbing Skripsi Berbasis *Tf-Idf* Dan *Cosine Similarity*

DOI: http://dx.doi.org/10.35889/jutisi.v15i1.3525

Creative Commons License 4.0 (CC BY – NC)

Adani Dharmawati[1*], Tri Wahyu Qur'ana[2], Rezky Izzatul Yazidah Anwar[3]
Teknik Informatika, UNISKA MAB, Banjarmasin, Indonesia
**e-mail Corresponding Author*: *adani.dharmawati@gmail.com*

Abstract

The selection of undergraduate thesis supervisors is a critical academic process; however, it is often conducted manually and lacks systematic consideration of topic–expertise alignment. This study proposes a thesis supervisor recommendation system based on text modeling techniques, employing Term Frequency–Inverse Document Frequency (TF-IDF) for term weighting and cosine similarity to measure relevance between students' thesis titles and lecturers' scientific publications. The system generates the top three recommended supervisors and is evaluated by directly comparing the results with faculty-assigned supervisors, achieving a matching rate of 12.22 percent. Although the matching rate is relatively low, the recommendations demonstrate strong academic relevance. Unlike previous studies, this work evaluates recommendation outcomes against institutional assignments, positioning the system as a decision support tool rather than a replacement mechanism. The findings indicate that the proposed approach has the potential to enhance efficiency, transparency, and objectivity in the thesis supervisor selection process.

Keywords: *Recommendation system; Thesis supervisor; Undergraduate thesis; TF-IDF; Cosine similarity*

Abstrak

Pemilihan dosen pembimbing skripsi merupakan proses akademik yang krusial, namun pada banyak perguruan tinggi masih dilakukan secara manual dan belum secara sistematis mempertimbangkan kesesuaian topik dengan keahlian dosen. Penelitian ini mengusulkan sistem rekomendasi dosen pembimbing berbasis pemodelan teks dengan menggunakan *Term Frequency–Inverse Document Frequency* (TF-IDF) sebagai metode pembobotan dan cosine similarity untuk mengukur relevansi antara judul skripsi mahasiswa dan publikasi ilmiah dosen. Sistem menghasilkan tiga rekomendasi dosen teratas dan dievaluasi dengan membandingkannya secara langsung terhadap dosen pembimbing yang ditetapkan oleh fakultas, dengan tingkat kecocokan sebesar 12,22 persen. Meskipun nilai kecocokan relatif rendah, rekomendasi yang dihasilkan menunjukkan relevansi akademik yang baik. Berbeda dengan penelitian sebelumnya, penelitian ini memosisikan sistem sebagai alat pendukung pengambilan keputusan, bukan sebagai pengganti penetapan institusional. Hasil penelitian menunjukkan potensi sistem dalam meningkatkan efisiensi, transparansi, dan objektivitas pemilihan dosen pembimbing skripsi.

Kata kunci: *Sistem rekomendasi; Dosen pembimbing; Skripsi; TF-IDF; Cosine similarity*

1. Pendahuluan

Pemilihan dosen pembimbing skripsi merupakan tahapan krusial dalam proses akademik mahasiswa karena berpengaruh langsung terhadap arah penelitian, kualitas bimbingan, dan keberhasilan penyelesaian tugas akhir. Berbagai studi menunjukkan bahwa keselarasan antara topik penelitian mahasiswa dan bidang keahlian dosen pembimbing berkontribusi signifikan terhadap efektivitas bimbingan dan mutu skripsi yang dihasilkan [1]. Oleh karena itu, penentuan dosen pembimbing seharusnya tidak dipandang semata sebagai proses administratif, melainkan sebagai keputusan akademik yang strategis.

Namun, dalam praktiknya, penetapan dosen pembimbing di banyak perguruan tinggi masih dilakukan secara manual dengan mempertimbangkan ketersediaan dosen, pemerataan beban bimbingan, dan kebijakan internal fakultas. Pendekatan ini sering kali belum secara optimal memperhatikan kesesuaian antara topik penelitian mahasiswa dan kompetensi keilmuan dosen, sehingga mahasiswa mengalami kesulitan memperoleh pembimbing yang relevan dengan bidang penelitiannya [2]. Kondisi tersebut berpotensi menurunkan efektivitas proses bimbingan dan memperpanjang masa studi. Hal ini menunjukkan perlunya pendekatan yang lebih sistematis, objektif, dan berbasis data dalam mendukung proses penentuan dosen pembimbing skripsi [3].

Sejumlah penelitian terdahulu telah mengusulkan sistem rekomendasi dosen pembimbing berbasis komputasi untuk mengatasi permasalahan tersebut. Pendekatan yang digunakan meliputi *text mining* dengan algoritma *Naive Bayes Classifier* [4], *content-based filtering* berbasis TF-IDF dan *cosine similarity* [5], metode *information retrieval* seperti BM25 [6], hingga pemodelan topik menggunakan *Latent Dirichlet Allocation* (LDA) [3]. Selain itu, model bahasa berbasis *deep learning* seperti BERT juga telah diterapkan untuk meningkatkan representasi semantik teks dalam sistem rekomendasi akademik [7]. Meskipun pendekatan-pendekatan tersebut menunjukkan kinerja teknis yang menjanjikan, sebagian besar penelitian masih berfokus pada akurasi rekomendasi secara algoritmik dan belum secara eksplisit mengevaluasi kesesuaian hasil rekomendasi terhadap penetapan dosen pembimbing oleh institusi.

Merujuk pada gap penelitian tersebut, studi ini mengembangkan model rekomendasi pembimbing skripsi berbasis analisis kemiripan topik melalui integrasi teknik pembobotan TF-IDF dan pengukuran kedekatan vektor menggunakan *cosine similarity*. Sistem diterapkan pada data judul skripsi mahasiswa dan publikasi ilmiah dosen di lingkungan internal fakultas. Kebaruan penelitian ini terletak pada evaluasi hasil rekomendasi dengan membandingkannya secara langsung terhadap dosen pembimbing yang ditetapkan oleh fakultas, sehingga sistem diposisikan sebagai alat pendukung pengambilan keputusan akademik, bukan sebagai pengganti mekanisme penetapan institusional.

2. Metodologi

2.1 Tahapan Penelitian

Penelitian ini mengikuti alur metodologis sistem rekomendasi berbasis kemiripan teks yang terdiri atas beberapa tahap berurutan. Tahap awal dilakukan studi literatur untuk mengidentifikasi kerangka kerja sistem rekomendasi berbasis konten, metode representasi teks, serta metrik evaluasi yang umum digunakan. Metode pembobotan *Term Frequency–Inverse Document Frequency* (TF-IDF) dan *cosine similarity* dipilih karena telah terbukti efektif dalam merepresentasikan dan membandingkan dokumen teks pada berbagai aplikasi sistem rekomendasi berbasis konten [8].

Tahap selanjutnya adalah identifikasi permasalahan dalam proses penentuan dosen pembimbing skripsi, yaitu adanya potensi ketidaksesuaian antara topik penelitian mahasiswa dan bidang keahlian dosen. Proses pengumpulan data mencakup pengambilan judul skripsi mahasiswa sebagai representasi kebutuhan penelitian serta kumpulan judul publikasi ilmiah dosen sebagai representasi profil keilmuan. Data teks kemudian dinormalisasi melalui proses praproses yang terdiri atas *cleaning*, *case folding*, *tokenization*, *stopword filtering*, dan *stemming* untuk mengurangi unsur tidak relevan serta memastikan uniformitas representasi fitur, sebagaimana standar implementasi pada model rekomendasi berbasis pemrosesan teks. [9].

Dataset pada studi ini terdiri dari 90 judul skripsi mahasiswa serta 48 dosen, dengan akumulasi 919 judul publikasi ilmiah dosen yang dikoleksi tanpa pembatasan rentang waktu publikasi. Pada tahap pembobotan teks, TF-IDF diterapkan menggunakan parameter *minimum document frequency* (*min_df*) sebesar 1, tanpa penerapan *maximum document frequency* (*max_df*), serta menggunakan daftar *stopword* bawaan (*default stopword list*) sesuai dengan bahasa masing-masing dokumen.

Setelah *preprocessing*, setiap dokumen direpresentasikan sebagai vektor fitur menggunakan skema TF-IDF yang mencerminkan tingkat kepentingan relatif setiap *term* dalam korpus. Vektor TF-IDF ini kemudian digunakan untuk menghitung tingkat kesamaan antara judul skripsi mahasiswa dan profil publikasi dosen menggunakan cosine similarity, di mana nilai *cosine* merepresentasikan kedekatan semantik antar dokumen dalam ruang fitur teks.

Pendekatan serupa telah digunakan dalam sistem rekomendasi berbasis TF-IDF dan *cosine similarity* pada berbagai domain [10].

Berdasarkan skor *cosine similarity* yang dihasilkan, sistem menyusun peringkat dosen untuk setiap judul skripsi mahasiswa dan menghasilkan tiga rekomendasi teratas (*Top-3 recommendation*) sebagai keluaran sistem. Tahap evaluasi dilakukan dengan menggunakan metrik *Hit@K* dan *Mean Reciprocal Rank* (MRR), dengan membandingkan hasil rekomendasi sistem terhadap dosen pembimbing yang ditetapkan oleh fakultas. Skema evaluasi ini digunakan untuk menilai kualitas peringkat rekomendasi dalam konteks sistem rekomendasi berbasis konten dan telah diterapkan pada penelitian sistem rekomendasi teks sebelumnya [11].

2.2 *Text Preprocessing*

Tahapan prapemrosesan teks merupakan bagian krusial dalam proses analisis data tekstual karena berfungsi mentransformasikan data mentah menjadi format yang siap untuk tahap komputasi selanjutnya. Proses ini bertujuan meminimalkan gangguan (noise), menyeragamkan struktur penulisan, serta memastikan bahwa perhitungan tingkat kemiripan antar dokumen dapat dilakukan secara lebih presisi dan konsisten. Dalam berbagai penelitian yang memanfaatkan metode Term Frequency–Inverse Document Frequency (TF-IDF) dan *cosine similarity* sebagai teknik representasi serta pengukuran kesamaan dokumen, tahap praproses memegang peranan penting. Rangkaian langkah seperti *case folding*, *tokenization*, *stopword filtering*, dan *stemming* terbukti mampu meningkatkan kualitas representasi fitur teks sebelum dilakukan pembobotan dan evaluasi similarity. Sebagai contoh, studi mengenai identifikasi kemiripan dokumen berbasis TF-IDF dan *cosine similarity* juga menerapkan beberapa tahapan normalisasi teks untuk mengoptimalkan hasil analisis. [12].

Langkah-langkah yang dilakukan dalam tahap pra-pemrosesan teks dalam penelitian ini terdiri dari beberapa fase sebagai berikut:

1) *Cleaning*, yaitu proses menghapus karakter khusus, angka, URL, emoji, serta spasi berlebih agar teks hanya berisi kata-kata yang memiliki makna relevan terhadap topik penelitian. Tahap ini bertujuan untuk mengurangi noise yang dapat memengaruhi proses analisis teks selanjutnya.
2) *Case Folding*, yaitu merupakan tahapan normalisasi teks yang dilakukan dengan mentransformasikan semua huruf menjadi bentuk lower-case untuk memastikan konsistensi representasi kata serta mengeliminasi perbedaan akibat variasi kapitalisasi.
3) *Tokenization*, yaitu merupakan proses pemecahan teks menjadi unit kata atau token sebagai dasar pengolahan lebih lanjut, sehingga masing-masing kata dapat direpresentasikan dan diberi bobot berdasarkan jumlah kemunculannya
4) *Stopword Removal*, yaitu merupakan tahapan penyaringan dengan menghilangkan kosakata berfrekuensi tinggi namun rendah nilai informatifnya, agar representasi teks lebih menitikberatkan pada token yang relevan secara semantik.
5) *Stemming*, yaitu merupakan proses reduksi morfologis yang bertujuan mengembalikan kata berimbuhan ke bentuk dasar melalui bantuan pustaka seperti Sastrawi, agar kata-kata dengan akar serupa direpresentasikan secara konsisten dan tidak menghasilkan bobot yang berbeda.

Penggunaan langkah-langkah tersebut dalam *text preprocessing* merupakan praktik umum dalam penelitian berbasis TF-IDF dan *cosine similarity* untuk tugas seperti klasifikasi teks, penilaian *similarity*, atau rekomendasi konten, dan merupakan bagian penting agar representasi fitur teks lebih efektif serta terbebas dari informasi yang tidak relevan. Integrasi TF-IDF dan *cosine similarity* biasanya dilengkapi *preprocessing* untuk meningkatkan kualitas fitur teks [12].

Setelah seluruh proses praproses diselesaikan, setiap dokumen selanjutnya direpresentasikan dalam bentuk vektor fitur dengan memanfaatkan metode pembobotan TF-IDF. Komponen *term frequency* (TF) merefleksikan jumlah kemunculan suatu istilah pada dokumen tertentu, sedangkan *inverse document frequency* (IDF) menggambarkan tingkat signifikansi istilah tersebut di dalam keseluruhan kumpulan dokumen. Dengan mekanisme ini, kata-kata yang sering muncul secara umum akan memperoleh bobot lebih kecil dibandingkan istilah yang lebih jarang namun memiliki nilai informatif yang lebih tinggi. Secara matematis, bobot TF-IDF dihitung sebagai berikut:

$$TF\text{-}IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{df(t)}\right) \quad (1)$$

di mana nilai **TF-IDF (t,d)** merepresentasikan tingkat kemunculan istilah *t* pada dokumen tertentu, **df(t)** menunjukkan banyaknya dokumen yang memuat istilah *t*, dan **N** menyatakan total keseluruhan dokumen yang terdapat dalam korpus.

Vektor hasil pembobotan TF-IDF kemudian dimanfaatkan sebagai dasar evaluasi kemiripan dokumen dengan pendekatan *cosine similarity*, yaitu ukuran matematis yang merepresentasikan kedekatan dua vektor berdasarkan sudut yang terbentuk di dalam ruang berdimensi TF-IDF. Rentang nilai *cosine similarity* berkisar antara 0 hingga 1; semakin besar nilainya, semakin kecil sudut yang terbentuk antar vektor, yang berarti representasi semantik kedua dokumen semakin berdekatan dan kemiripan kontennya semakin kuat. Metode ini banyak digunakan dalam sistem rekomendasi berbasis teks dan sistem temu kembali informasi untuk menilai relevansi antar dokumen teks berbobot TF-IDF. *Cosine similarity* diukur dari sudut antar-vektor TF-IDF untuk menghitung tingkat kesamaan dokumen [2].

2.3 Evaluasi

Evaluasi sistem rekomendasi dosen pembimbing dilakukan untuk memastikan model yang dibangun tidak hanya menghasilkan daftar peringkat berdasarkan kesamaan topik, tetapi juga memberikan rekomendasi yang relevan secara empiris. Karena hasil keluaran berupa peringkat dosen untuk setiap judul skripsi, maka digunakan metrik evaluasi berbasis *ranking* yang banyak dipakai dalam penelitian sistem rekomendasi konten dan *information retrieval*, yaitu Hit@K dan *Mean Reciprocal Rank* (MRR). Metrik-metrik ini sering digunakan dalam studi rekomendasi konten untuk menilai kemampuan model dalam menempatkan item relevan dalam daftar rekomendasi teratas. Salah satunya adalah Hit@K dan MRR yang didefinisikan dalam literatur evaluasi sistem rekomendasi [13].

Hit@K digunakan untuk menghitung proporsi rekomendasi di mana dosen pembimbing aktual muncul pada posisi urutan sampai K pertama dalam daftar rekomendasi. Secara formal, *Hit@K* untuk satu mahasiswa dapat dirumuskan sebagai:

$$Hit@K = \frac{1}{|U|}\sum_{u=1}^{|U|} I(rank_u \leq K) \quad (2)$$

di mana *N* menyatakan banyaknya mahasiswa dalam dataset uji, r*ank*$_u$ merepresentasikan posisi dosen pembimbing aktual bagi mahasiswa ke-*u* pada daftar rekomendasi yang dihasilkan, sedangkan *I(.)* merupakan fungsi indikator yang menghasilkan nilai 1 ketika syarat terpenuhi dan 0 dalam kondisi sebaliknya. *Hit@K* memberikan gambaran proporsi hasil di mana item relevan (dosen aktual) berhasil ditempatkan di dalam *top-K*, sehingga semakin tinggi nilai *Hit@K*, semakin sering sistem menempatkan rekomendasi relevan pada posisi atas.

Sementara itu, *Mean Reciprocal Rank* (MRR) mengukur rata-rata kebalikan dari posisi munculnya pertama dosen pembimbing aktual dalam daftar rekomendasi. Secara matematis, *MRR* dapat dinyatakan sebagai:

$$MRR = \frac{1}{|U|}\sum_{u=1}^{|U|}\frac{1}{rank_u} \quad (3)$$

di mana *rank*$_u$ adalah posisi pertama dosen pembimbing aktual dalam daftar peringkat untuk mahasiswa ke-u. *MRR* memberikan bobot lebih besar pada rekomendasi yang relevan muncul lebih tinggi dalam daftar, karena nilai kebalikan peringkat akan semakin dekat ke 1 jika item relevan berada di posisi atas. Metrik ini penting dalam sistem rekomendasi ranking karena memperhitungkan urutan hasil rekomendasi, tidak hanya keberadaannya. MRR merupakan rata-rata dari kebalikan posisi peringkat pertama item relevan dalam daftar rekomendasi [13].

Dengan demikian, kombinasi Hit@K dan MRR memberikan gambaran kinerja sistem baik dari segi keberhasilan menempatkan item relevan dalam daftar atas (*Hit@K*), maupun

kemampuan sistem untuk meletakkan item relevan sedekat mungkin dengan urutan pertama (*MRR*), yang sangat relevan dalam konteks rekomendasi berbasis peringkat.

3. Hasil dan Pembahasan

3.1 Pengumpulan Data

Tahap awal penelitian dilakukan dengan mengumpulkan file berisi judul skripsi mahasiswa serta file berisi daftar publikasi atau judul penelitian dari setiap dosen. Data skripsi tidak hanya mencakup judul penelitian mahasiswa, tetapi juga memuat informasi mengenai dosen pembimbing yang telah ditetapkan oleh fakultas. Data pembimbing asli inilah yang kemudian dijadikan sebagai acuan (*ground truth*) dalam proses evaluasi, sehingga hasil rekomendasi sistem dapat dibandingkan secara langsung dengan kondisi nyata. Sementara itu, daftar publikasi dosen digunakan untuk membangun profil keilmuan masing-masing dosen, yang menjadi dasar perhitungan kesesuaian topik menggunakan metode TF-IDF dan *cosine similarity*.

Tabel 1. Sampel Data Publikasi Dosen

Nama Dosen	Judul Publikasi
Dosen A	*A comparison of root architecture and shoot morphology of Cercis siliquastrum L. between two water regimes*
Dosen B	*A Comprehensive Evaluation of Machine Learning Classifiers in Named Entity Recognition*
Dosen C	*A preliminary model analysis of knowledge management design: spiritual leadership on knowledge worker productivity*
Dosen D	*AI Implementation Impact on Workforce Productivity: The Role of AI Training and Organizational Adaptation*
Dosen E	Air mancur otomatis dengan musik berbasis arduino

Tabel 2. Sampel Data Judul Proposal Skripsi Mahasiswa

Nama Mahasiswa	Judul Skripsi
Mahasiswa A	Sistem Informasi Manajemen Keuangan Dan Penggajian Pada Cv Peninsula Abadi Berbasis Web
Mahasiswa B	Sistem Informasi Manajemen Pengajuan Satyalancana Karya Satya (Slks) Bagi Pns Di Kabupaten Tabalong Melalui Simpeg Tabalong Berbasis Web
Mahasiswa C	Sistem Informasi Permintaan Darah Dan Pendaftaran Donor Darah Unit Transfusi Darah (Utd) Rsud Ulin Banjarmasin Berbasis Web
Mahasiswa D	Sistem Informasi Pelanggan Dan Suplai Air Pada Pt Air Minum Murakata Lestari (Perseroda) Berbasis Web
Mahasiswa E	Sistem Informasi Pengelolaan Buku Tamu Dan Antrian Digital Berbasis Web Dengan Integrasi *Qr-Code* Pada Disdukcapil Banjarmasin

3.2 Prapengolahan Data Teks

Proses prapengolahan data teks dilakukan melalui beberapa tahapan agar data yang digunakan lebih bersih dan konsisten. Tahap pertama adalah normalisasi, yaitu menghapus karakter khusus maupun angka yang tidak relevan serta mengubah seluruh teks menjadi huruf kecil (*lowercase*) agar seragam. Selanjutnya dilakukan tokenisasi, yakni memecah teks menjadi potongan kata (token) untuk memudahkan analisis lebih lanjut. Setelah itu, kata-kata umum yang tidak memiliki makna penting, atau dikenal sebagai *stopword*, dihapus menggunakan daftar *stopword* standar Bahasa Indonesia. Tahapan berikutnya adalah *stemming* atau *lemmatization* dengan bantuan pustaka seperti Sastrawi, yang berfungsi mengembalikan kata ke bentuk dasarnya sehingga variasi kata tidak mengganggu proses perhitungan. Terakhir, seluruh judul publikasi milik masing-masing dosen digabungkan menjadi satu dokumen profil keilmuan dosen yang utuh, yang kemudian digunakan sebagai representasi utama dalam perhitungan kesesuaian dengan judul skripsi mahasiswa.

Tabel 3. Tahapan Prapengolahan Data Teks

Tahapan	Deskripsi	Tujuan	Contoh
Normalisasi	Menghapus karakter khusus dan mengubah huruf menjadi kecil	Menyamakan format teks	“Kinerja!” → “kinerja”
Tokenisasi	Memecah teks menjadi kata	Memudahkan analisis kata	“kinerja mahasiswa” → [“kinerja”, “mahasiswa”]
Stopword Removal	Menghapus kata umum yang tidak bermakna	Mengurangi *noise*	“yang”, “dan”, “di”
Stemming	Mengembalikan kata ke bentuk dasar	Menyatukan variasi kata	“mempengaruhi” → “pengaruh”
Penggabungan	Menggabungkan semua publikasi dosen	Membentuk profil keilmuan dosen	Dokumen dosen A = judul1 + judul2 + ...

3.3 Representasi Fitur (TF-IDF)

Setelah proses prapemrosesan teks diselesaikan, seluruh dokumen kemudian dikonversi ke dalam format numerik dengan menerapkan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). Transformasi ini dilakukan untuk merepresentasikan bobot atau tingkat signifikansi suatu istilah dalam satu dokumen dibandingkan dengan keseluruhan kumpulan dokumen. Dalam penelitian ini, korpus yang dianalisis bersumber dari dua jenis data utama, yakni judul skripsi mahasiswa serta judul karya ilmiah dosen.

Secara konseptual, komponen *Term Frequency* (TF) merefleksikan jumlah kemunculan sebuah istilah pada dokumen tertentu, sementara *Inverse Document Frequency* (IDF) menunjukkan tingkat kelangkaan istilah tersebut di dalam seluruh dokumen pada korpus, sehingga kata yang jarang muncul akan memperoleh bobot yang relatif lebih besar dibandingkan kata yang sering digunakan secara umum.

Dalam implementasinya, model TF-IDF dibangun menggunakan skema *unigram* untuk merepresentasikan kata tunggal, dengan ambang batas frekuensi minimum (*min_df*) yang disesuaikan dengan ukuran korpus guna mengurangi pengaruh kata yang terlalu jarang muncul. Normalisasi vektor menggunakan metode L2 diterapkan untuk menstandarkan panjang vektor dokumen, sehingga perhitungan kesamaan antarvektor menjadi lebih konsisten dan dapat dibandingkan secara adil.

Sebagai ilustrasi mekanisme pembobotan, Tabel 4 menyajikan contoh nilai TF, IDF, dan TF-IDF dari beberapa kata yang muncul pada korpus dosen dan mahasiswa. Contoh ini bertujuan untuk memperjelas proses representasi fitur, bukan sebagai hasil akhir perhitungan sistem.

Tabel 4. Contoh Bobot TF-IDF pada Korpus Dosen dan Mahasiswa

Kata	TF	IDF	TF-IDF	Keterangan
analisis	3	1.2	3.6	Kata umum namun sering muncul
machine	2	2.1	4.2	Kata spesifik bidang AI
learning	2	2.1	4.2	Berkorelasi dengan “machine”
pendidikan	1	1.8	1.8	Umum di bidang sosial
biomedis	1	3.0	3.0	Kata jarang dan sangat spesifik

3.4 Perhitungan Kesamaan

Pengukuran tingkat kesamaan dilakukan untuk mengetahui derajat kedekatan antara judul skripsi mahasiswa dan profil publikasi dosen yang telah direpresentasikan dalam bentuk vektor berbasis TF-IDF. Pendekatan yang diterapkan dalam penelitian ini adalah *cosine similarity*, yaitu metode yang mengevaluasi kemiripan dua vektor dengan mempertimbangkan sudut yang terbentuk di antara keduanya. Semakin kecil sudut antar vektor, semakin tinggi skor kesamaan yang diperoleh, sehingga menunjukkan relevansi topik dosen yang semakin sesuai dengan judul skripsi mahasiswa.

Secara matematis, nilai *cosine similarity* untuk dua vektor A dan Bdapat dinyatakan dengan rumus berikut:

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

dengan *A* merepresentasikan vektor TF-IDF dari judul skripsi mahasiswa, sedangkan *B*adalah vektor TF-IDF yang menggambarkan profil publikasi dosen. Notasi *A.B* menunjukkan hasil operasi perkalian titik (*dot product*) antara kedua vektor tersebut, sementara ||A| dan ||B|| masing-masing menyatakan panjang atau norma dari setiap vektor.

Proses perhitungan kesamaan dilakukan dengan membandingkan satu judul skripsi mahasiswa terhadap seluruh vektor profil publikasi dosen dalam korpus. Untuk setiap dosen, sistem menghitung nilai *cosine similarity* antara vektor judul mahasiswa dan vektor dosen yang bersangkutan. Nilai kesamaan yang dihasilkan berada pada rentang 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan tingkat kemiripan yang lebih besar.

Setelah seluruh nilai *cosine similarity* diperoleh, dosen diurutkan berdasarkan skor kesamaan dari nilai tertinggi hingga terendah. Berdasarkan hasil pengurutan tersebut, sistem memilih tiga dosen dengan nilai kesamaan tertinggi untuk disajikan sebagai rekomendasi dosen pembimbing skripsi. Tahapan pengurutan dan pemilihan rekomendasi ini bertujuan untuk memberikan alternatif pembimbing yang paling relevan secara akademik terhadap topik penelitian mahasiswa.

Tabel 5. Contoh Nilai Cosine Similarity

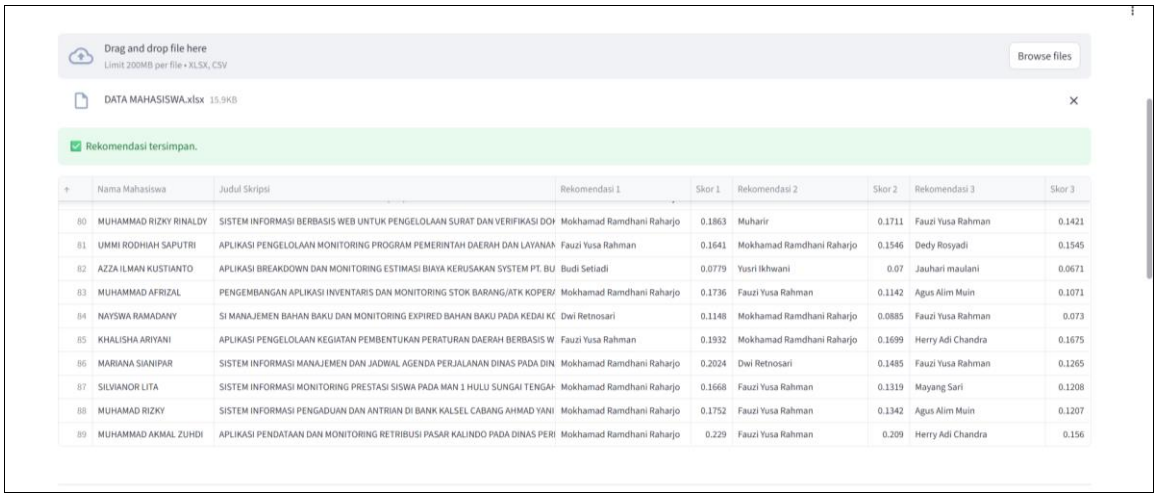
Nama Dosen	Nilai Cosine Similarity	Peringkat	Status Rekomendasi
Dosen A	0.87	1	Direkomendasikan
Dosen B	0.82	2	Direkomendasikan
Dosen C	0.79	3	Direkomendasikan
Dosen D	0.54	4	Tidak direkomendasikan

Status rekomendasi ditentukan berdasarkan hasil pengurutan nilai *cosine similarity*. Dosen dengan tiga nilai *cosine similarity* tertinggi ditetapkan sebagai Direkomendasikan, sesuai dengan tujuan sistem yang menghasilkan tiga alternatif dosen pembimbing yang paling relevan dengan topik skripsi mahasiswa. Dosen dengan peringkat di luar tiga teratas dikategorikan sebagai Tidak Direkomendasikan. Pendekatan ini digunakan untuk memastikan konsistensi dan objektivitas dalam proses pemilihan dosen pembimbing.

3.5 Implementasi Prototipe

Pengembangan prototipe dilakukan dengan memanfaatkan skrip *Python* yang dijalankan di *Google Colab* sebagai *backend* untuk memproses input judul skripsi mahasiswa, melakukan transformasi TF-IDF, serta menghitung *cosine similarity* dengan profil publikasi dosen. Untuk memudahkan interaksi pengguna, dibuat antarmuka sederhana menggunakan *Streamlit*, yang memungkinkan mahasiswa memasukkan judul skripsi secara langsung melalui web dan secara otomatis menampilkan Top-3 rekomendasi dosen berdasarkan skor *similarity* tertinggi. Prototipe ini mengintegrasikan logika pemrosesan teks dan perhitungan kesamaan dengan tampilan visual yang intuitif, sehingga proses pengujian dan penggunaan sistem menjadi lebih praktis tanpa memerlukan pengaturan lingkungan pemrograman yang kompleks.

Gambar 1. Tampilan Aplikasi Prototipe Dengan Streamlit



	Nama Mahasiswa	Judul Skripsi	Rekomendasi 1	Skor 1	Rekomendasi 2	Skor 2	Rekomendasi 3	Skor 3
80	MUHAMMAD RIZKY RINALDY	SISTEM INFORMASI BERBASIS WEB UNTUK PENGELOLAAN SURAT DAN VERIFIKASI DOK	Mokhamad Ramdhani Raharjo	0.1863	Muharir	0.1711	Fauzi Yusa Rahman	0.1421
81	UMMI RODHIAH SAPUTRI	APLIKASI PENGELOLAAN MONITORING PROGRAM PEMERINTAH DAERAH DAN LAYANAN	Fauzi Yusa Rahman	0.1641	Mokhamad Ramdhani Raharjo	0.1546	Dedy Rosyadi	0.1545
82	AZZA ILMAN KUSTIANTO	APLIKASI BREAKDOWN DAN MONITORING ESTIMASI BIAYA KERUSAKAN SYSTEM PT. BU	Budi Setiadi	0.0779	Yusri Ikhwani	0.07	Jauhari maulani	0.0671
83	MUHAMMAD AFRIZAL	PENGEMBANGAN APLIKASI INVENTARIS DAN MONITORING STOK BARANG/ATK KOPERA	Mokhamad Ramdhani Raharjo	0.1736	Fauzi Yusa Rahman	0.1142	Agus Alim Muin	0.1071
84	NAYSWA RAMADANY	SI MANAJEMEN BAHAN BAKU DAN MONITORING EXPIRED BAHAN BAKU PADA KEDAI KO	Dwi Retnosari	0.1148	Mokhamad Ramdhani Raharjo	0.0885	Fauzi Yusa Rahman	0.073
85	KHALISHA ARIYANI	APLIKASI PENGELOLAAN KEGIATAN PEMBENTUKAN PERATURAN DAERAH BERBASIS W	Fauzi Yusa Rahman	0.1932	Mokhamad Ramdhani Raharjo	0.1699	Herry Adi Chandra	0.1675
86	MARIANA SIANIPAR	SISTEM INFORMASI MANAJEMEN DAN JADWAL AGENDA PERJALANAN DINAS PADA DIN	Mokhamad Ramdhani Raharjo	0.2024	Dwi Retnosari	0.1485	Fauzi Yusa Rahman	0.1265
87	SILVIANOR LITA	SISTEM INFORMASI MONITORING PRESTASI SISWA PADA MAN 1 HULU SUNGAI TENGAH	Mokhamad Ramdhani Raharjo	0.1668	Fauzi Yusa Rahman	0.1319	Mayang Sari	0.1208
88	MUHAMAD RIZKY	SISTEM INFORMASI PENGADUAN DAN ANTRIAN DI BANK KALSEL CABANG AHMAD YANI	Mokhamad Ramdhani Raharjo	0.1752	Fauzi Yusa Rahman	0.1342	Agus Alim Muin	0.1207
89	MUHAMMAD AKMAL ZUHDI	APLIKASI PENDATAAN DAN MONITORING RETRIBUSI PASAR KALINDO PADA DINAS PERI	Mokhamad Ramdhani Raharjo	0.229	Fauzi Yusa Rahman	0.209	Herry Adi Chandra	0.156

Gambar 2. Tampilan Halaman Hasil Perhitungan

3.6 Evaluasi

Proses evaluasi dilakukan dengan membandingkan hasil rekomendasi sistem terhadap dosen pembimbing yang ditetapkan oleh fakultas. Alur evaluasi dapat dijelaskan sebagai berikut:

1) Sistem menerima input berupa judul dan abstrak tugas akhir mahasiswa.
2) Data teks mahasiswa dan dosen direpresentasikan menggunakan metode TF-IDF.
3) Tingkat kemiripan antara topik penelitian mahasiswa dan keilmuan dosen dihitung menggunakan *cosine similarity*.
4) Sistem menghasilkan daftar dosen pembimbing dalam bentuk peringkat (ranking) berdasarkan nilai kemiripan tertinggi.
5) Hasil rekomendasi tersebut dibandingkan dengan dosen pembimbing aktual (*ground truth*) yang ditetapkan fakultas.

Untuk mengukur performa sistem, digunakan beberapa metrik evaluasi yang umum pada sistem rekomendasi, yaitu:

1) Hit@1, untuk mengukur apakah dosen pembimbing fakultas muncul pada peringkat pertama rekomendasi.
2) Hit@3, untuk mengukur apakah dosen pembimbing fakultas muncul dalam tiga rekomendasi teratas.
3) *Mean Reciprocal Rank* (MRR), untuk mengukur posisi rata-rata dosen pembimbing fakultas dalam daftar rekomendasi.

Hasil evaluasi performa sistem rekomendasi dipaparkan pada Tabel 6.

Tabel 6. Hasil Evaluasi

Metrik	Nilai (%)
Hit@1	2,22
Hit@2	7,78
Hit@3	12,22
MRR	6,48

Hasil evaluasi menunjukkan bahwa nilai Hit@1 dan *Mean Reciprocal Rank* (MRR) yang diperoleh relatif rendah, yang mengindikasikan bahwa dosen pembimbing yang ditetapkan oleh fakultas jarang muncul pada peringkat teratas rekomendasi sistem. Hal ini menunjukkan bahwa sistem belum sepenuhnya mereplikasi pola penetapan pembimbing yang diterapkan oleh fakultas.

Namun, nilai Hit@3 sebesar 12,22 persen menunjukkan bahwa dalam sejumlah kasus, dosen pembimbing yang ditetapkan fakultas masih termasuk dalam tiga rekomendasi teratas yang dihasilkan sistem. Temuan ini menunjukkan bahwa sistem mampu mengidentifikasi relevansi topik keilmuan antara judul skripsi mahasiswa dan profil publikasi dosen, meskipun belum secara konsisten menempatkan dosen pembimbing fakultas pada peringkat pertama.

3.7 Pembahasan

Hasil evaluasi menunjukkan bahwa sistem rekomendasi yang diusulkan belum sepenuhnya mampu mereplikasi pola penetapan dosen pembimbing oleh fakultas. Rendahnya nilai Hit@1 dan MRR mengindikasikan bahwa dosen pembimbing yang ditetapkan fakultas jarang muncul pada peringkat teratas rekomendasi sistem, meskipun nilai Hit@3 sebesar 12,22% menunjukkan bahwa dalam sebagian kasus dosen tersebut masih termasuk dalam tiga rekomendasi teratas yang dihasilkan sistem.

Temuan ini menunjukkan bahwa pendekatan pemilihan dosen pembimbing berbasis kesesuaian topik ilmiah melalui pemodelan teks memiliki potensi untuk mendukung proses penentuan pembimbing, namun belum sepenuhnya mencerminkan kebijakan fakultas. Hal ini disebabkan karena penetapan pembimbing di tingkat institusi tidak hanya mempertimbangkan kesesuaian topik penelitian, tetapi juga berbagai faktor non-akademik, seperti pemerataan beban kerja dosen, ketersediaan waktu bimbingan, pengalaman dosen, serta kebijakan administratif internal [2].

Temuan penelitian ini memperlihatkan keselarasan dengan studi-studi sebelumnya. yang menyatakan bahwa sistem rekomendasi berbasis konten cenderung menghasilkan tingkat kecocokan yang terbatas ketika ground truth dipengaruhi oleh faktor subjektif atau kebijakan non-teknis [6]. Penelitian oleh Damayanti *et al.* [1] dan Wijanto *et al.* [2] menunjukkan bahwa pendekatan TF-IDF dan *cosine similarity* lebih merepresentasikan kedekatan topik ilmiah dibandingkan keputusan manual institusi. Sementara itu, penelitian lain yang mengombinasikan pendekatan berbasis konten dengan aturan atau kebijakan institusi melaporkan peningkatan akurasi rekomendasi, meskipun dengan konsekuensi peningkatan kompleksitas sistem [14].

Dengan demikian, penelitian ini menegaskan bahwa pendekatan berbasis kesamaan teks efektif dalam mengidentifikasi relevansi keilmuan [15], namun belum memadai untuk sepenuhnya menggantikan proses penugasan pembimbing yang bersifat administratif. Kontribusi utama penelitian ini terletak pada penyediaan sistem rekomendasi yang objektif dan transparan sebagai *decision support system*, yang berfungsi sebagai alat bantu awal dalam memberikan alternatif pembimbing yang relevan secara akademik dan selanjutnya dapat dipertimbangkan bersama faktor kebijakan institusional lainnya.

4. Simpulan

Penelitian ini berhasil mengembangkan aplikasi rekomendasi dosen pembimbing skripsi berbasis metode TF-IDF dan *cosine similarity* untuk mendukung proses pemilihan pembimbing yang lebih objektif, efisien, dan relevan secara akademik. Hasil evaluasi menunjukkan bahwa tingkat kesesuaian dengan penetapan fakultas masih relatif rendah akibat pertimbangan administratif, namun sistem mampu memberikan rekomendasi alternatif yang transparan dan berfokus pada keselarasan bidang keilmuan antara mahasiswa dan dosen. Temuan ini menunjukkan potensi pendekatan analisis teks dalam meningkatkan efektivitas dan akuntabilitas proses akademik di perguruan tinggi. Pengembangan selanjutnya disarankan

mengadopsi metode berbasis semantik, seperti *word embeddings* atau *model transformer*, serta melakukan standarisasi dan perluasan data publikasi dosen. Integrasi sistem dengan kebijakan fakultas juga diperlukan agar implementasi aplikasi dapat memberikan manfaat optimal bagi mahasiswa dan institusi.

Daftar Referensi

[1] A. Damayanti, F. Purwani and M. Kadafi, "A Content-Based Thesis Supervisor Recommendation System Based on Research Interest Clustering and Cosine Similarity," *JUSIFO (Jurnal Sistem Informasi),* vol. 2, pp. 111-120, 2025.

[2] M. C. Wijanto, R. Rachmadiany and O. Karnalim, "Thesis Supervisor Recommendation with Representative Content and Information Retrieval," *Journal ofInformation Systems Engineeringand Business Intelligence,* vol. 2, no. 2, pp. 143-150, 2020.

[3] L. R. Nisa, A. Luthfiarta, A. Nugraha, M. Hasan, A. Wulandari and A. . M. Huda, "A Topic-Based Approach For Recommending Undergraduate Thesis Supervisor Using Lda With Cosine Similarity," *Jurnal Teknik Informatika (JUTIF),* vol. 6, no. 1, pp. 311-323, 2025.

[4] M. Ikhsan and R. K. R, "Penerapan Text Mining pada Sistem Rekomendasi Pembimbing Skripsi Mahasiswa Menggunakan Algoritma Naïve Bayes Classifier," *Indonesian Journal of Computer Science,* vol. 12, no. 6, pp. 4196-4207, 2023.

[5] F. D. Astuti and W. Andriyani, "Development of a Final Project Supervisor Recommendation System Using," *JIKO (Jurnal Informatika Dan Komputer),* vol. 9, no. 2, pp. 474-483, 2025.

[6] A. A. B. Arisetiawan, I. and D. E. Ratnawati, "Sistem Rekomendasi Dosen Pembimbing Berdasarkan Dokumen Judul," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 3, no. 6, pp. 5832-5836, 2019.

[7] F. T. Sabilillah, S. Winarno and R. . B. Abiyyi, "Implementasi BERT dan Cosine Similarity untuk Rekomendasi Dosen Pembimbing berdasarkan Judul Tugas Akhir," *Edumatic: Jurnal Pendidikan Informatika /,* vol. 8, no. 2, p. 585–594, 2024.

[8] F. A. Hidayat and L. Fitriani, "Aplikasi Mentor Pembelajaran BerbasisSistem Rekomendasi Content-Based Filteringdengan Metode TF-IDF dan Cosine Similarity," *Jurnal Algoritma,* vol. 22, no. 2, pp. 1048-1059, 2025.

[9] M. Farhan, M. R. Andreawan, R. Antonius and N. D. Ulhag, "Evaluasi Dan Pengembangan Sistem Rekomendasi Game Berbasis Content-Based Filtering Dengan TF-IDF Dan Cosine Similarity," *BINER : Jurnal Ilmu Komputer, Teknik dan Multimedia,* vol. 2, no. 4, pp. 400-4008, 2024.

[10] S. Oyadila, D. Abdullah and A. R. , "Implementasi Content-Based Filtering Dengan Tf-Idf Dan Cosine Similarity Untuk Sistem Rekomendasi Destinasi Wisata Di Aceh Tengah," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab,* vol. 10, no. 2, p. 1329–1339, 2025.

[11] I. K. Phan and Y. , "Perancangan Skema Evaluasi untuk Sistem Rekomendasi Berita," *TIN: Terapan Informatika Nusantara,* vol. 6, no. 8, pp. 1351-1360, 2026.

[12] A. H. Nasrullah, "Integrasi Tf-Idf Dan Algoritma Cosine Similarity Untuk Deteksi Tingkat Kemiripan Judul Penelitian (Studi Kasus Mahasiswa Fakultas Ilmu Komputer UNISAN Gorontalo)," *Information Technology Education Journal,* vol. 3, no. 3, p. 113–118, 2024.

[13] D. Ç. Ertuğrul and S. Bitirim, "Job recommender systems: a systematic literature review, applications, open issues, and challenges," *Journal of Big Data,* vol. 12, no. 1, p. art. 140, 2025.

[14] M. . A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. . I. Mulloh and H. A. Gozali, "Improving Text Preprocessing For Student Complaint," *IOP Conference Series: Materials Science and Engineering,* pp. 1-6, 2019.

[15] A. Widianto, E. Pebriyanto, F. and M. , "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.