Efektivitas TF-IDF dan Cosine Similarity untuk Deteksi Plagiarisme pada Judul Skripsi

Syauqi Zainun Nauval*[1], Fidi Supriadi[2], David Setiadi[3]
[1,2]Universitas Sebelas April; Jl. Angkrek Situ No.19, Kec. Sumedang Utara, Kabupaten Sumedang, Jawa Barat 45323
[3]Program Studi Informatika, Fakultas Teknologi Informasi, Sumedang
e-mail: ***[1]zainunnauval@gmail.com**, [2]fidi@unsap.ac.id, [3]david@unsap.ac.id

Abstrak

Deteksi kemiripan judul skripsi merupakan langkah krusial untuk mencegah plagiarisme dan menjaga integritas akademik di institusi pendidikan. Penelitian ini bertujuan untuk mengevaluasi efektivitas gabungan metode Term Frequency–Inverse Document Frequency (TF-IDF) dan Cosine Similarity dalam mengukur derajat kemiripan antarjudul skripsi. Pendekatan ini dipilih karena TF-IDF mampu menghasilkan representasi vektor numerik dengan membobotkan kata kunci yang signifikan dalam dokumen, dan Cosine Similarity terbukti akurat dalam mengukur kemiripan berdasarkan orientasi vektor. Dataset yang digunakan terdiri dari 1.917 judul skripsi unik mahasiswa yang diperoleh dari repositori Institut Pertanian Bogor (IPB). Proses analisis meliputi tahapan prapemrosesan teks (seperti case folding, tokenizing, stopword removal, dan stemming), pembobotan kata menggunakan TF-IDF, dan pengukuran kemiripan menggunakan Cosine Similarity. Hasil eksperimen menunjukkan bahwa metode ini berhasil mengidentifikasi 35 pasangan judul dengan tingkat kemiripan tinggi (nilai cosine similarity > 0.8) dan secara efektif membedakan lebih dari 1,8 juta pasangan judul lainnya yang bersifat unik. Oleh karena itu, kombinasi TF-IDF dan Cosine Similarity dinilai sangat efektif untuk mendeteksi potensi plagiarisme pada teks pendek seperti judul skripsi, menjadikannya alat preventif yang kuat.

Kata kunci— Deteksi Plagiarisme, Text Similarity, TF-IDF, Cosine Similarity, Judul Skripsi

Abstract

The detection of thesis title similarity is a critical and essential step in preventing plagiarism and maintaining academic integrity within educational institutions. This study aims to evaluate the effectiveness of combining the Term Frequency–Inverse Document Frequency (TF-IDF) and Cosine Similarity methods in measuring the degree of resemblance between thesis titles. This approach is chosen because TF-IDF successfully generates numerical vector representations by weighting unique and significant keywords in the documents, and Cosine Similarity has proven accurate in measuring similarity based on vector orientation. The dataset utilized consists of 1,917 unique student thesis titles obtained from the Bogor Agricultural University (IPB) repository. The analysis process includes rigorous text preprocessing steps (such as case folding, tokenizing, stopword removal, and stemming), word weighting using TF-IDF, and similarity measurement via Cosine Similarity. The experimental results show that this combined method successfully identified 35 pairs of titles with a high degree of similarity (a cosine similarity score > 0.8) and effectively differentiated over 1.8 million other unique title pairs. Consequently, the combination of TF-IDF and Cosine Similarity is considered highly effective for detecting potential plagiarism in short texts like thesis titles, establishing it as a strong preventative tool in academic environments.

Keywords— Plagiarism Detection, Text Similarity, TF-IDF, Cosine Similarity, Thesis Titles

1. PENDAHULUAN

Integritas akademik di lembaga pendidikan tinggi senantiasa terancam oleh praktik plagiarisme, yang dikategorikan sebagai pelanggaran serius dan dapat dikenai sanksi tegas [1]. Plagiarisme sendiri diartikan sebagai tindakan peniruan atau penyalinan hasil karya orang lain tanpa menyertakan kredit sumber aslinya. Fenomena ini diperparah oleh pesatnya perkembangan teknologi informasi, terutama dengan hadirnya *Artificial Intelligence* (AI) generatif, yang memfasilitasi kemudahan dalam melakukan penjiplakan [2]. Kondisi yang ada mewajibkan institusi akademik untuk mengimplementasikan sistem pendeteksian dan pencegahan yang canggih. Salah satu upaya preventif yang dinilai efektif adalah pelaksanaan pemeriksaan kemiripan sejak tahap awal pengajuan judul skripsi atau tugas akhir [3]**.** Meskipun perbandingan judul tergolong analisis *short-text similarity*, adanya kemiripan judul dapat menjadi indikasi kuat terhadap kesamaan topik, yang berpotensi menyebabkan duplikasi penelitian secara keseluruhan, sehingga langkah ini sangat krusial.

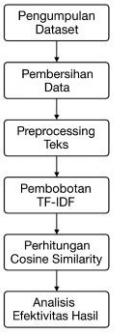
Kebutuhan akan sistem deteksi yang mutakhir menjadi mendesak untuk membendung plagiarisme, dimulai dari proses awal pengajuan judul. Walaupun judul hanya berupa teks pendek, kesamaan antarjudul dapat menjadi petunjuk adanya kesamaan tema riset secara menyeluruh [4]**.** Dengan demikian, pemeriksaan kemiripan judul skripsi dapat dioptimalkan sebagai langkah antisipatif sebelum proses penulisan formal dimulai.

Dalam analisis kemiripan teks, Vector Space Model (VSM) adalah kerangka kerja populer yang merepresentasikan dokumen sebagai vektor numerik [5]. VSM mengandalkan skema pembobotan kata menggunakan algoritma Term Frequency–Inverse Document Frequency (TF-IDF**)** [6]. Bobot TF-IDF akan meningkat secara signifikan untuk kata-kata kunci yang spesifik pada suatu dokumen tetapi jarang ditemukan dalam koleksi dokumen secara keseluruhan, sehingga efektif menonjolkan kata kunci yang relevan. Setelah pembobotan, **Cosine Similarity** digunakan untuk menghitung tingkat kemiripan antarvektor judul [7]**.** Cosine Similarity sering dipilih karena kemampuannya dalam mengukur kesamaan berdasarkan orientasi vektor, memberikan hasil pencocokan yang lebih presisi dalam konteks penemuan informasi (*information retrieval)* [8]. Secara terpisah maupun gabungan, penerapan TF-IDF dan Cosine Similarity telah teruji keefektifannya dalam *clustering* dan klasifikasi dokumen berbasis teks [9].

Kombinasi TF-IDF dan Cosine Similarity sendiri telah terbukti berhasil dalam berbagai kasus yang relevan, termasuk deteksi plagiarisme skripsi [10]. Namun, masih terdapat celah penelitian (*research gap*) terkait validasi spesifik metode ini pada himpunan data judul skripsi dalam volume besar, serta kebutuhan untuk menetapkan ambang batas (*threshold*) kemiripan yang paling optimal untuk jenis teks pendek [11]. Berangkat dari urgensi ini, penelitian yang kami lakukan berfokus untuk menganalisis dan mengevaluasi efektivitas gabungan metode Term Frequency–Inverse Document Frequency (TF-IDF) dan Cosine Similarity dalam mengukur tingkat kemiripan antarjudul skripsi, sebagai kontribusi nyata dalam upaya mendeteksi potensi plagiarisme di lingkungan akademik.

2. METODE PENELITIAN

Penelitian ini mengadopsi pendekatan kuantitatif eksperimental yang dirancang untuk menganalisis serta menguji tingkat efektivitas gabungan metode Term Frequency–Inverse Document Frequency (TF-IDF) dan Cosine Similarity dalam pengukuran derajat kemiripan antarjudul skripsi [1]. Seluruh aktivitas riset ini dilaksanakan di bawah kerangka kerja **Vector** Space Model (VSM). Prosesnya mencakup serangkaian tahapan inti yang terstruktur, yaitu: pengumpulan dan pembersihan data, dilanjutkan dengan prapemrosesan teks, kemudian pembobotan menggunakan algoritma TF-IDF, diikuti oleh perhitungan Cosine Similarity, dan diakhiri dengan analisis efektivitas hasil temuan [7]. Alur lengkap pelaksanaan penelitian ini diringkas sebagaimana ditampilkan pada Gambar 1.



Gambar 1. Alur Tahapan Penelitian

2. 1 Pengumpulan Data

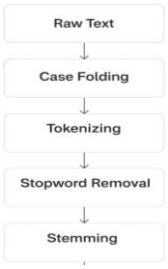
Data riset untuk penelitian ini didapatkan dari platform Kaggle, yang merupakan hasil kompilasi data yang bersumber dari repositori skripsi Institut Pertanian Bogor (IPB). Berkas data yang digunakan adalah format *Microsoft Excel (.xlsx)*. File tersebut memiliki beberapa bidang informasi, termasuk kolom Judul (*Title*), Abstrak berbahasa Indonesia, dan Abstrak berbahasa Inggris. Namun, fokus analisis kami hanya menggunakan kolom judul skripsi berbahasa Indonesia saja, sejalan dengan tujuan utama studi untuk mengevaluasi kemiripan antarjudul skripsi. Himpunan data ini terdiri dari ratusan judul skripsi dari beragam fakultas di IPB. Judul-judul ini dipilih karena karakternya sebagai teks pendek (*short-text*) yang sangat relevan untuk menguji efektivitas metode TF-IDF dan Cosine Similarity [12] [13].

2. 2 Pembersihan Data

Sebelum pemrosesan lebih lanjut, dilakukan tahap pembersihan data yang esensial untuk menjamin kualitas dan validitas himpunan data[7]. Kegiatan *cleaning* ini melibatkan beberapa langkah kunci: eliminasi entri judul yang terduplikasi, penghapusan data yang hilang (*missing values*), serta penghilangan semua karakter non-alfabet seperti simbol, tanda baca, dan angka [5]. Proses pembersihan ini wajib dilakukan untuk memastikan data yang akan digunakan pada tahap selanjutnya murni, valid, dan siap untuk analisis komputasional.

2. 3 Prapemrosesan Teks

Tahap Prapemrosesan Teks merupakan langkah fundamental untuk menstandarisasi seluruh data teks, menjadikannya optimal dan siap untuk pemrosesan komputasional [14]. *Proses ini melibatkan beberapa tahapan wajib, yaitu: Case Folding (penyamaan kasus huruf), Tokenizing (pemecahan kata), Stopword Removal (penghapusan kata umum), dan Stemming (pencarian kata dasar)* [6]. Rincian alur kerja dari proses prapemrosesan teks ini dapat dilihat secara detail pada Gambar 2.



Gambar 2. Alur Detail Tahapan Prapemrosesan Teks

2. 4 Pembobotan Term Frequency–Inverse Document Frequency (TF-IDF)

Term Frequency–Inverse Document Frequency (TF-IDF) adalah skema pembobotan yang diterapkan untuk mengonversi data teks menjadi representasi numerik (vektor) berdasarkan tingkat kepentingan setiap kata. Algoritma ini dirancang untuk meningkatkan fokus pada kata-kata kunci yang memiliki nilai signifikan dalam satu dokumen, jika dibandingkan dengan frekuensi kemunculannya di keseluruhan koleksi dokumen [10], [11]. Dengan kata lain, TF-IDF berfungsi memodelkan teks sebagai vektor numerik, di mana bobot untuk setiap kata (*term*) dihitung menggunakan kombinasi rumus berikut:

$$TF(t,d) = \frac{f(t,d)}{\sum_{k \in d} f(k,d)} \quad (1)$$

$$IDF(t) = \log \frac{N}{1+n(t)} \quad (2)$$

$$W_{t,d} = TF_{t,d} \times IDF_t \quad (3)$$

Keterangan:

- $W_{t,d}$: Bobot akhir kata (*term*) t dalam dokumen d.
- $f(t,d)$: Frekuensi kemunculan kata t dalam dokumen d (Judul Skripsi).
- $\sum_{k \in d} f(k,d)$: Jumlah total kata dalam dokumen d.
- N : Jumlah total dokumen (Judul Skripsi) dalam dataset.
- $n(t)$: Jumlah dokumen yang mengandung kata t.

2. 5 Perhitungan Cosine Similarity

Setelah proses pembobotan kata menggunakan TF-IDF selesai, langkah selanjutnya dalam penelitian ini adalah menghitung tingkat kemiripan antara setiap pasangan judul skripsi yang kini telah direpresentasikan sebagai vektor numerik [1], [7], [11]. Metode yang digunakan untuk tujuan ini adalah Cosine Similarity. Algoritma ini bekerja dengan mengukur besar sudut yang terbentuk di antara dua vektor; skor 1 menunjukkan kesamaan yang identik (kemiripan sempurna), sedangkan skor 0 menunjukkan ketiadaan kemiripan. Dengan demikian, tingkat kemiripan antarjudul dihitung melalui Cosine Similarity menggunakan rumus matematis sebagai berikut:

$$Cosine\ Similarity(A,B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4)$$

Keterangan:

- $A\ dan\ B$: Vektor bobot TF-IDF dari dokumen $A\ dan\ B$.
- $A_i\ dan\ B_i$: Kompenen bobot ke- i dari vektor $A\ dan\ B$.
- n : Jumlah kata unik yang ada dalam kedua dokumen.

2.6 Analisis Efektivitas

Tahap akhir ini melibatkan analisis mendalam untuk mengevaluasi seberapa efektif kombinasi metode ini dalam mengidentifikasi potensi plagiarisme [1] [7]. Untuk mempermudah interpretasi, skor kemiripan yang diperoleh diklasifikasikan ke dalam tiga kategori utama, yang didasarkan pada ambang batas (*threshold*) yang telah ditetapkan:

- Tinggi: Skor Cosine Similiarity lebih dari 0.8, mengindikasikan adanya kesamaan substansial.
- Sedang: Skor Cosine Similarity berkisar antara 0.5 hingga 0.8, menunjukkan adanya kemiripan sebagian kata kunci.

- Rendah: Skor Cosine Similarity kurang dari 0.5, menandakan bahwa kedua judul bersifat unik atau sangat berbeda.

Keseluruhan temuan ini kemudian dieksplorasi untuk menentukan kapabilitas gabungan TF-IDF dan Cosine Similarity dalam membedakan seacara akurat antara judul-judul yang benar-benar mirip dan yang tidak mirip.

3. HASIL DAN PEMBAHASAN

Temuan-temuan yang diperoleh dari implementasi metode TF-IDF dan Cosine Similarity terhadap himpunan data judul skripsi. Selain itu, bagian ini mencakup diskusi mendalam terkait tingkat efektivitas metode tersebut dalam mengidentifikasi kemiripan teks. Seluruh temuan eksperimen disajikan dengan memanfaatkan tabel dan grafik yang berfungsi untuk memvisualisasikan hasil dan mendukung analisis keefektifan metode.

3.1 Pengumpulan dan Pembersihan Data

Data primer yang dimanfaatkan dalam penelitian ini bersumber dari repository skripsi Institut Pertanian Bogor (IPB) yang dapat diakases oleh publik melalui platform Kaggle. Data mentah tersebut mengandung koleksi judul dan abstrak skripsi dari berbagai fakultas. Meskipun demikian, analisis kemiripan utama hanya berfokus pada kolom judul skripsi saja. Sebelum data di proses lebih lanjut, tahap pembersihan data dilaksanakan seacara ketat untuk menjamin konsistensi dan kualitas himpunan data. Langkah-langkah detail yang diambil dalam tahap pembersihan adalah:

- Eliminasi judul yang terduplikasi (judul skripsi yang identik) untuk mencegah bias saat pengukuran kemiripan.
- Penghilangan entri yang kosong (missing values) yang tidak memiliki informasi pada kolom judul.
- Pembersihan karakter non-alfabet, termasuk angka, simbol, dan tanda baca yang tidak relevan dengan analisis teks.

Data awal yang terkumpul berjumlah 1.958 judul. Setelah melalui proses penghapusan data duplikat dan entri kosong, jumlah data unik yang tersisa dan siap diolah adalah 1917 judul. Hasil pembersihan data ini dapat dilihat secara ringkas pada Tabel 1.

Tabel 1. Hasil Pembersihan Data

Jumlah data awal	**Jumlah data setelah di bersihkan**
1.958	1.917

Hasil dari tahap pembersihan data tersebut menjamin bahwa data judul yang akan diproses adalah valid dan relevan. Data yang sudah dimurnikan ini kemudian siap digunakan untuk tahap berikutnya, yaitu prapemrosesan teks dan selanjutnya adalah pembobotan TF-IDF dalam analisis kemiripan.

3.2 Prapemrosesan Teks

Himpunan data yang diperoleh pada dasarnya masih berupa data mentah atau belum terstruktur. Guna menunjang implementasi efektif dari metode TF-IDF dan Cosine Similarity, seluruh data teks harus melalui serangkaian prapemrosesan teks. Proses ini dilakasanakan dengan memanfaatkan perangkat Google Colab dan bahasa pemrograman Python. Tujuan utama dari prapemrosesan teks adalah menyeragamkan format data teks sehingga memadai untuk dianalisis secara komputansional.

3.2.1 Case Folding

Langkah permulaan dalam prapemrosesan teks adalah Case Folding. Ini adalah prosedur pengubahan seluruh karakter huruf dalam teks menjadi format huruf kecil (*lowercase*). Prosedur ini krusial agar sistem tidak membedakan antara kapitalisasi huruf yang memiliki arti yang sama. Selain itu, *case folding* juga berfungsi untuk mengeliminasi karakter-karakter yang bukan alfabet, seperti tanda baca, simbol, dan angka yang tidak relevan. Hasil dari proses case folding dapat diamati pada Gambar 3.

Sebelum Case Folding	Setelah Case Folding
Peramalan Nilai Ekspor Migas Indonesia Menggunakan Algoritma Long ShortTerm Memory	peramalan nilai ekspor migas indonesia menggunakan algoritma long shortterm memory
Implementasi LSTM AutoencoderOCSVM dalam Pendeteksian Anomali Cuaca untuk Pertanian Studi Kasus Jawa Timur	implementasi lstm autoencoderocsvm dalam pendeteksian anomali cuaca untuk pertanian studi kasus jawa timur
Perbandingan Performa Long ShortTerm Memory dan XGBoost dalam Memprediksi Curah Hujan Harian di Sumatera Selatan	perbandingan performa long shortterm memory dan xgboost dalam memprediksi curah hujan harian di sumatera selatan

Gambar 3. Proses Case Folding pada Data Judul Skripsi

Pemeriksaan hasil *case folding* menunjukkan bahwa semua karakter huruf telah berhasil diubah menjadi huruf kecil (*lowercase*), dan karakter-karakter yang tidak relevan (non-alfabet) telah dihilangkan.

3.2.2 Tokenizing

Tokenizing adalah langkah kedua dalam proses text preprocessing. Tahap ini didefinisikan sebagai prosedur pemisahan atau pemecahan kalimat menjadi unit-unit kata yang berdiri sendiri. Tujuan utama dari tokenizing adalah agar setiap kata dalam jdudl skripsi dapat dianalisis secara indenpenden dalam proses analisis teks selanjutnya. Proses pemecahan ini dilakukan dengan memanggil fungsi tokenizing yang tersedia dalam pustaka NLTK (Natural Languange Toolkit) di lingkungan pemrograman Python. Hasil dari proses tokenizing ini disajikan pada Gambar 4.

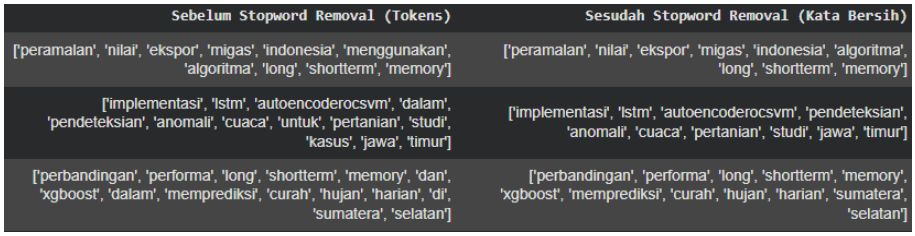
Sebelum Tokenizing	Sesudah Tokenizing (Tokens)
peramalan nilai ekspor migas indonesia menggunakan algoritma long shortterm memory	['peramalan', 'nilai', 'ekspor', 'migas', 'indonesia', 'menggunakan', 'algoritma', 'long', 'shortterm', 'memory']
implementasi lstm autoencoderocsvm dalam pendeteksian anomali cuaca untuk pertanian studi kasus jawa timur	['implementasi', 'lstm', 'autoencoderocsvm', 'dalam', 'pendeteksian', 'anomali', 'cuaca', 'untuk', 'pertanian', 'studi', 'kasus', 'jawa', 'timur']
perbandingan performa long shortterm memory dan xgboost dalam memprediksi curah hujan harian di sumatera selatan	['perbandingan', 'performa', 'long', 'shortterm', 'memory', 'dan', 'xgboost', 'dalam', 'memprediksi', 'curah', 'hujan', 'harian', 'di', 'sumatera', 'selatan']

Gambar 4. Proses Tokenizing pada Data Judul Skripsi

Hasil yang diperoleh dari proses tokenisasi menunjukkan bahwa setiap judul skripsi telah berhasil dipecah menjadi kumpulan kata-kata tunggal (*token*). Melalui pemisahan ini, setiap kata kini dapat diidentifikasi secara individual oleh sistem sebagai unit dasar untuk analisis.

3.2.3 Stopword Removal

Langkah berikutnya dalam prapemrosesan adalah penghilang Stopword (Stopword Removal). Ini merupakan prosedur untuk menghapus kata-kata frekuensi tinggi yang secara linguistik umum, namun minin kontribusi makna dalam konteks analisis teks. Oleh karena itu, proses eliminasi ini dijalankan dengan memanfaatkan daftar stopword standar Bahasa Indonesia yang tersedia di pustaka Sastrawi dalam bahasa pemrograman Python. Hasil dari proses penghapusan stopword disajikan pada Gambar 5.



Sebelum Stopword Removal (Tokens)	Sesudah Stopword Removal (Kata Bersih)
['peramalan', 'nilai', 'ekspor', 'migas', 'indonesia', 'menggunakan', 'algoritma', 'long', 'shortterm', 'memory']	['peramalan', 'nilai', 'ekspor', 'migas', 'indonesia', 'algoritma', 'long', 'shortterm', 'memory']
['implementasi', 'lstm', 'autoencoderocsvm', 'dalam', 'pendeteksian', 'anomali', 'cuaca', 'untuk', 'pertanian', 'studi', 'kasus', 'jawa', 'timur']	['implementasi', 'lstm', 'autoencoderocsvm', 'pendeteksian', 'anomali', 'cuaca', 'pertanian', 'studi', 'jawa', 'timur']
['perbandingan', 'performa', 'long', 'shortterm', 'memory', 'dan', 'xgboost', 'dalam', 'memprediksi', 'curah', 'hujan', 'harian', 'di', 'sumatera', 'selatan']	['perbandingan', 'performa', 'long', 'shortterm', 'memory', 'xgboost', 'memprediksi', 'curah', 'hujan', 'harian', 'sumatera', 'selatan']

Gambar 5. Hasil Proses Stopword Removal pada Judul Skripsi

Hasil dari proses ini menunjukkan bahwa kata-kata fungsional yang tidak relevan telah sukses dieliminasi, menyisakan kata-kata kunci yang substantif untuk digunakan dalam analisis berikutnya.

3.2.4 Stemming

Langkah pamungkas dalam rangkaian text preprocessing adalah Stemming. Stemming adalah prosedur untuk mereduksi setiap kata menjadi bentuk dasar atau kata akar (root word) aslinya. Fungksi utama dari stemming adalah menyeragamkan berbagai bentuk turunan kata yang memiliki arti serupa, sehingga memastikan konsistensi data dalam analisis teks. Proses stemming ini diimplementasikan menggunakan pustaka Sastrawi di python. Pustaka ini secara khusus dirancang untuk Bahasa Indonesia, memungkinkannya mengidentifikasi dan menghilangkan imbuhan (awalan, akhiran, dan sisipan) secara efektif. Hasil konkret dari proses stemming dapat diamati pada Gambar 6.

Sebelum Stemming (Kata Bersih)	Sesudah Stemming (Kata Dasar)
['peramalan', 'nilai', 'ekspor', 'migas', 'indonesia', 'algoritma', 'long', 'shortterm', 'memory']	['amal', 'nilai', 'ekspor', 'migas', 'indonesia', 'algoritma', 'long', 'shortterm', 'memory']
['implementasi', 'lstm', 'autoencoderocsvm', 'pendeteksian', 'anomali', 'cuaca', 'pertanian', 'studi', 'jawa', 'timur']	['implementasi', 'lstm', 'autoencoderocsvm', 'deteksi', 'anomali', 'cuaca', 'tani', 'studi', 'jawa', 'timur']
['perbandingan', 'performa', 'long', 'shortterm', 'memory', 'xgboost', 'memprediksi', 'curah', 'hujan', 'harian', 'sumatera', 'selatan']	['banding', 'performa', 'long', 'shortterm', 'memory', 'xgboost', 'prediksi', 'curah', 'hujan', 'hari', 'sumatera', 'selatan']

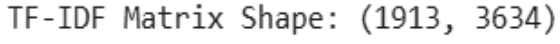
Gambar 6. Hasil Proses Stemming pada Judul Skripsi

Setelah tahap stemming selesai, seluruh teks telah berada dalam bentuk standar yang siap digunakan untuk tahap selanjutnya, yaitu pembobotan kata menggunakan metode TF-IDF. Dengan demikian, proses *text preprocessing* ini memastikan bahwa data teks yang digunakan sudah bersih, terstruktur, dan konsisten, sehingga hasil perhitungan kemiripan antarjudul dapat lebih akurat dan relevan.

3.3 Pembobotan Term Frequency–Inverse Document Frequency (TF-IDF)

Langkah berikutnya setelah teks melewati semua tahapan preprocessing adalah pemberian bobot kata melalui algoritma Term Frequency-Inverse Document Frequency (TF-IDF). Metode ini berperan krusial dalam mentransformasi data teks menjadi representasi numerik (vektor), yang nilai-nilainya mencerminkan tingkat kepentingan setiap kata kunci di dalam sebuah dokumen.

Dalam model ini, setiap judul skripsi di modelkan sebagai titik dalam ruang vektor, ditentukan oleh bobot kata-kata yang terkandung di dalamnya. Logika dasarnya adalah jika suatu kata sering muncul dalam judul tertentu tetapi jarang ditemukan di judul-judul lain dalam koleksi, maka bobot TF-IDF-nya akan semakin tinggi, menandakan kekhasan kata tersebut. Visualisasi dari bobot kata yang dihasilkan ini dapat diamati dalam bentuk matriks TF-IDF pada Gambar 7.



Gambar 7. Hasil matriks TF-IDF

Hal ini membuktikan bahwa setiap judul skripsi direpresentasikan oleh 3.634 bobot numerik yang menunjukkan tingkat kepentingan setiap kata dalam dokumen tersebut. contoh sebagian kecil nilai bobot TF-IDF ditunjukkan pada Tabel 2.

Tabel 2. Contoh Hasil Pembobotan TF-IDF pada Lima Judul Skripsi

Kata	1	2	3	4	5
algoritma	0.3046	0.0000	0.0000	0.0000	0.2777
amal	0.2544	0.0000	0.0000	0.0000	0.0000
analisis	0.0000	0.0000	0.0000	0.1268	0.1268
anomali	0.0000	0.3813	0.0000	0.0000	0.0000
autoencoderocsvm	0.0000	0.4181	0.0000	0.0000	0.0000

Berdasarkan tabel tersebut, terlihat bahwa kata **"autoencoderocsvm"** memiliki bobot tertinggi pada Judul 2 karena kata tersebut jarang muncul pada judul lainnya. Sementara kata **"algoritma"** memiliki bobot di beberapa judul karena sering digunakan dalam berbagai topik penelitian.

3.4 Perhitungan Cosine Similarity

Setelah proses pembobotan kata selesai dilakukan dengan TF-IDF, langkah berikutnya dalam alur kerja adalah mengukur tingkat kemiripan anatara judul-judul skripsi yang telah diubah menjadi vektor numerik, yaitu menggunakan metode Cosine Similarity. Cosine Similarity menghitung besar sudut antara dua vektor TF-IDF yang mewakili setiap judul skripsi. Semakin kecil sudutnya, semakin mirip kedua teks tersebut. Nilai Cosine Similarity berkisar antara 0 hingga 1, dengan:

- 1 menunjukkan dua judul identik,
- 0 menunjukkan tidak ada kemiripan.

Berdasarkan hasil perhitungan pada dataset skripsi Institut Pertanian Bogor (IPB), diperoleh 35 pasangan judul dengan nilai kemiripan di atas 0.8 (kategori tinggi). 10 (Sepuluh) contoh pasangan judul dengan tingkat kemiripan tertinggi ditampilkan pada Tabel 3.

Tabel 3. Contoh Hasil Perhitungan Cosine Similarity

NO	Judul 1	Judul 2	Similarity
1	Peramalan Inflow Bulanan DAS Saguling Berdasarkan Curah Hujan Bulanan	Peramalan inflow bulanan DAS Saguling berdasarkan curah hujan bulanan	1.000
2	Uji desain baru produk kembang gula X menggunakan perancangan percobaan	Uji Desain Baru Produk Kembang Gula X Menggunakan Perancangan Percobaan	1.000
3	Perbandingan regresi fungsi pangkat dan regresi polinom ordo tiga untuk menduga model sebaran data yang tidak setangkup dan mempunyai satu lengkungan	Perbandingan Regresi Fungsi Pangkat dan Regresi Polinom Ordo Tiga Untuk Menduga Model Sebaran Data Yang Tidak Setangkup dan Mempunyai Satu Lengkungan	1.000
4	Analisis regresi logistik terhadap	Analisis Regresi	

	faktorfaktor yang mempengaruhi status gizi anak balita di Jawa Tengah	Logistik Terhadap FaktorFaktor Yang Mempengaruhi Status Gizi Anak Balita di Jawa Tengah	1.000
5	Penerapan model Kano pada analisis kepuasan pelanggan PT XYZ	Penerapan Model Kano Pada Analisis Kepuasan Pelanggan Pt Xyz	1.000

Hasil perhitungan menunjukkan bahwa seluruh pasangan dengan nilai kemiripan tertinggi (≥ 0,8) umumnya memiliki perbedaan minor seperti kapitalisasi huruf, tanda baca, atau variasi kecil dalam penulisan, bukan perbedaan substansi. Dngan ini membuktikan bahwa metode *TF-IDF* dan *Cosine Similarity* mampu mendeteksi kemiripan redaksional secara akurat.

3.5 Analisis efektivitas

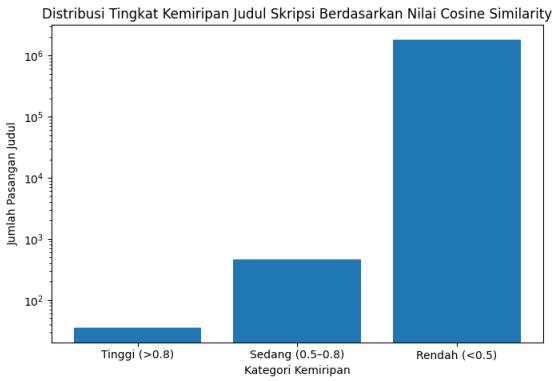
Analisis efektivitas dilakukan untuk mengevaluasi kemampuan metode *TF-IDF* dan *Cosine Similarity* dalam mendiferensiasi berbagai tingkat kemiripan antarjudul skripsi dievaluasi secara cermat. Untuk mempermudah interpretasi hasil, skor kemiripan yang dihasilkan dikelompokkan ke dalam tiga kategori utama, sebagaimana disajikan pada Gambar 8.

```
Jumlah pasangan dengan kemiripan tinggi (>0.8): 35
Jumlah pasangan dengan kemiripan sedang (0.5–0.8): 467
Jumlah pasangan dengan kemiripan rendah (<0.5): 1828326
```

Gambar 8**.** Klasifikasi Nilai Cosine Similarity

Dari hasil klasifikasi tersebut, terlihat bahwa mayoritas besar pasangan judul skripsi memiliki nilai kemiripan yang tergolong rendah, yaitu sebanyak 1.828.326 pasangan, yang menandakan bahwa mayoritas judul skripsi bersifat unik dan tidak menunjukkan indikasi kemiripan tinggi. Sementara itu, terdapat 35 pasangan dengan nilai kemiripan tinggi (> 0.8) yang menunjukkan adanya kemiripan struktur kalimat dan kata kunci yang kuat antarjudul. Sebagian lainnya, yaitu 467 pasangan, berada pada kategori kemiripan sedang (0.5–0.8) yang mengindikasikan bahwa kedua judul memiliki kesamaan sebagian istilah namun tetap berbeda secara konteks dan topik penelitian.

Visualisasi distribusi hasil kemiripan dapat dilihat pada Gambar 9, yang menggambarkan proporsi tiap kategori kemiripan. Dari gambar tersebut terlihat bahwa proporsi terbesar berada pada kategori kemiripan rendah, diikuti oleh kemiripan sedang, dan hanya sebagian kecil yang termasuk kategori kemiripan tinggi.



Gambar 9. Distribusi Kategori Kemiripan Judul

Secara keseluruhan, hasil analisis menunjukkan bahwa kombinasi metode TF-IDF dan Cosine Similarity efektif dalam mengukur tingkat kemiripan antarjudul skripsi dan mampu membedakan antara judul yang mirip dan tidak mirip secara akurat.

4. KESIMPULAN

Penelitian ini telah menganalisis efektivitas metode Term Frequency–Inverse Document Frequency (TF-IDF) dan Cosine Similarity dalam mendeteksi kemiripan antarjudul skripsi. Berdasarkan hasil eksperimen terhadap 1.917 data judul dari repositori IPB, kombinasi kedua metode ini mampu mengidentifikasi 35 pasangan judul dengan tingkat kemiripan tinggi (nilai cosine similarity > 0,8) serta membedakan mayoritas judul lain yang bersifat unik.

Hasil tersebut menunjukkan bahwa tahapan text preprocessing yang tepat meliputi case folding, tokenizing, stopword removal, dan stemming berkontribusi signifikan terhadap peningkatan akurasi pembobotan TF-IDF dan hasil perhitungan Cosine Similarity. Secara ringkas, kombinasi metode ini menunjukkan keefektifan dan akurasi yang tinggi dalam mendeteksi indikasi plagiarisme pada teks berformat pendek dalam konteks akademik.

Oleh karena itu, implementasi gabungan TF-IDF dan Cosine Similarity layak menjadi fondasi bagi pengembangan sistem otomatis untuk validasi awal judul skripsi pada tahap pengajuan judul skripsi di perguruan tinggi, guna mendukung integritas akademik dan mencegah duplikasi topik penelitian.

Selain itu, hasil penelitian ini memberikan kontribusi praktis bagi lembaga akademik dalam meningkatkan proses validasi judul skripsi secara objektif dan terukur. Metode ini juga paling efektif dalam mendeteksi kesamaan konsep atau tema penelitian, bukan sekadar kesamaan kata, sehingga relevan untuk diterapkan pada tahap awal penentuan topik skripsi.

5. SARAN

Saran-saran berikut ditujukan untuk penelitian lebih lanjut (*future works*) guna meningkatkan akurasi sistem dan memperluas cakupan deteksi kemiripan, berdasarkan temuan yang diperoleh:

- Perluasan Dataset dan Generalisasi: Disarankan untuk memperluas lingkup dataset yang digunakan agar mencakup repositori dari berbagai perguruan tinggi lainnya. Pengujian dengan data yang lebih heterogen akan membantu menguji kemampuan generalisasi metode TF-IDF dan Cosine Similarity pada berbagai konteks penulisan akademik.
- Penerapan Metode *Word Embedding*: Penelitian selanjutnya dapat mencoba membandingkan kinerja TF-IDF/Cosine Similarity dengan metode *word embedding* (seperti BERT atau Word2Vec). Metode ini dapat mengevaluasi apakah analisis semantik dapat lebih akurat dalam mendeteksi kemiripan tematik pada teks pendek seperti judul, karena TF-IDF memiliki keterbatasan dalam memahami relasi antar kata (*semantic meaning*).
- Pengembangan Sistem *Hybrid*: Disarankan untuk mengembangkan sistem deteksi kemiripan *hybrid* yang tidak hanya membandingkan teks judul, tetapi juga mengintegrasikan perbandingan abstrak dan kata kunci untuk memberikan nilai kemiripan yang lebih komprehensif, sehingga dapat membedakan antara plagiarisme yang disengaja dan kebetulan.
- Optimasi *Preprocessing*: Melakukan analisis lebih lanjut terhadap *threshold* dan efektivitas setiap tahap *preprocessing* (*Stemming* dan *Stopword Removal*) secara spesifik pada istilah-istilah yang sering muncul di lingkup topik IPB untuk meningkatkan presisi.

Melalui pengembangan-pengembangan tersebut, diharapkan sistem deteksi kemiripan teks berbasis TF-IDF dan *Cosine Similarity* dapat terus disempurnakan untuk

mendukung integritas akademik serta mencegah plagiarisme di lingkungan perguruan tinggi.

DAFTAR PUSTAKA

[1] M. Azmi, "Analisis Tingkat Plagiasi Dokumen Skripsi Dengan Metode Cosine Similarity Dan Pembobotan Tf-Idf," *Tek. Teknol. Inf. dan Multimed.*, vol. 2, no. 2, hal. 90–95, 2022, doi: 10.46764/teknimedia.v2i2.51.

[2] R. Khalida, A. Rahmandri, S. A. Matilda Magren, dan E. Nurmiati, "Etika Teknologi Informasi dalam Dunia Pendidikan: Tinjauan Literatur atas Penggunaan AI dan Isu Plagiarisme Akademik," *J. SAINTEKOM*, vol. 15, no. 2, hal. 222–234, 2025, doi: 10.33020/saintekom.v15i2.928.

[3] A. Subadri dan I. Pratama, "Sistem Deteksi Plagiarism Pada Judul Tugas Akhir Menggunakan Metode Rabin-Karp Berbasis Web," *Technol. J. Ilm.*, vol. 13, no. 4, hal. 306, 2022, doi: 10.31602/tji.v13i4.7786.

[4] F. Representation dan C. Similarity, "Document Similarity using Term Frequency-Inverse Document Frequency Representation and Cosine Similarity," vol. 4, no. 2, hal. 149–153, 2024.

[5] A. E. Budiman dan A. Widjaja, "Analisis Pengaruh Teks Preprocessing Terhadap Deteksi Plagiarisme Pada Dokumen Tugas Akhir," *J. Tek. Inform. dan Sist. Inf.*, vol. 6, no. 3, hal. 475–488, 2020, doi: 10.28932/jutisi.v6i3.2892.

[6] M. U. Albab, Y. K. P., dan M. N. Fawaiq, "Optimization of the Stemming Technique on Text Preprocessing President 3 Periods Topic," *J. Transform.*, vol. 20, no. 2, hal. 1–12, 2023, doi: 10.26623/transformatika.v20i2.5374.

[7] O. A. Resta, A. Aditya, dan F. E. Purwiantono, "Plagiarism Detection in Students' Theses Using The Cosine Similarity Method," *SinkrOn*, vol. 5, no. 2, hal. 305–313, 2021, doi: 10.33395/sinkron.v5i2.10909.

[8] J. P. Pamput, A. R. Muthmainnah, D. F. Surianto, dan N. Fadilah, "Perbandingan Cosine Similarity dan Weighted Jaccard Similarity dalam Pengembangan Mesin Pencari Perpustakaan Digital," *J. Inform. J. Pengemb. IT*, vol. 10, no. 4, hal. 907–919, 2025, doi: 10.30591/jpit.v10i4.8773.

[9] V. Meida Hersianty, E. Larasati Amalia, D. Puspitasari, dan D. Wahyu Wibowo, "Penerapan Algoritma Tf-Idf Dan Cosine Similarity Dalam Sistem Rekomendasi Lowongan Pekerjaan," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 1, hal. 1619–1625, 2025, doi: 10.36040/jati.v9i1.12406.

[10] D. Septiani dan I. Isabela, "Term Frequency Inverse Document Frequency (Tf-Idf) Analysis in Information Retrieval in Text Documents," *J. Sist. dan Teknol. Inf. Indones.*, vol. 1, no. 2, hal. 81–88, 2022.

[11] R. Al Rasyid dan D. H. U. Ningsih, "Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 8, no. 1, hal. 170–178, 2024, doi: 10.35870/jtik.v8i1.1416.

[12] R. R. Anugrah, "PENERAPAN COSINE SIMILARITY DAN PEMBOBOTAN TF-IDF UNTUK KLASIFIKASI PENGADUAN MASYARAKAT BERBASIS WEB (Studi Kasus : BAGWASSIDIK DITRESKRIMUM POLDA KALBAR)," *Coding J. Komput. dan Apl.*, vol. 11, no. 1, hal. 100, 2023, doi: 10.26418/coding.v11i1.55598.

[13] J. Wong, I. Sanu, dan H. Irsyad, "Implementasi TF-IDF, Cosine Similarity, dan Logistic Regression Pada Rekomendasi Buku Berdasarkan Mood Pembaca Dengan Data Oversampling," *Device J. Inf. Syst. Comput. Sci. Inf. Technol.*, vol. 6, no. 1, hal. 142–154, 2025, doi: 10.46576/device.v6i1.6499.

[14] H. Sari, G. Leonarde Ginting, dan T. Zebua, "Penerapan Algoritma Text Mining Dan Tf-Idf Untuk Pengelompokan Topik Skripsi," *Terap. Inform. Nusant.*, vol. 2, no. 7, hal.

414–432, 2021, [Daring]. Tersedia pada: https://ejurnal.seminar-id.com/index.php/tin