Infotekmesin
Vol.16, No.02, Juli 2025
p-ISSN: 2087-1627, e-ISSN: 2685-9858
DOI: 10.35970/infotekmesin.v16i2.2812, pp. 328-337

Kombinasi Algoritma TF-IDF dan *Weighted Dice Similarity* Untuk Pengukuran Kemiripan Judul Tugas Akhir

Santi Purwaningrum[1*], Agus Susanto[2], Annas Setiawan Prabowo [3]
[1, 2]Program Studi Teknologi Rekayasa Multimedia, Politeknik Negeri Cilacap
[3]Program Studi Teknik Informatika, Politeknik Negeri Cilacap
[1,2,3]Jln. Dr. Soetomo No.1 Karangcengis Sidakaya, Kabupaten Cilacap, 53212, Indonesia
E-mail: santi.purwaningrum@pnc.ac.id[1], agus.susanto@pnc.ac.id [2], annassetiawanp@gmail.com[3]

Info Naskah:
Naskah masuk: 11 Juni 2025
Direvisi: 2 Juli 2025
Diterima: 6 Juli 2025

Abstrak

Tingginya tingkat kemiripan judul tugas akhir mahasiswa menjadi isu penting dalam menjaga orisinalitas karya ilmiah di lingkungan perguruan tinggi. Penelitian ini bertujuan mengembangkan sistem pendeteksi kemiripan judul secara otomatis dengan menggabungkan algoritma *Term Frequency–Inverse Document Frequency* dan *Weighted Dice Similarity*. Metode TF-IDF digunakan untuk memberikan bobot pada kata-kata penting dalam judul, sedangkan *Weighted Dice Similarity* digunakan untuk mengukur tingkat kesamaan antar judul berdasarkan distribusi dan bobot kata-kata tersebut. Penelitian ini menggunakan data judul tugas akhir yang telah melalui proses anotasi manual sebagai *ground truth*. Proses analisis melibatkan tahapan *preprocessing*, pembobotan kata, dan perhitungan *similarity* antar judul. Hasil penelitian menunjukkan bahwa sistem mencapai akurasi sebesar 94%, presisi 66,67%, recall 81,3%, serta nilai *similarity* rata-rata dengan metode *Weighted Dice* sebesar 0,62. Meskipun nilai presisi tidak terlalu tinggi, kombinasi kedua metode dinilai efektif karena mampu mengidentifikasi kemiripan judul berdasarkan representasi semantik dan struktur leksikal secara bersamaan, yang tidak ditangkap hanya dengan metode pembobotan atau pengukuran kesamaan saja.

Keywords:
term frequency–inverse document frequency;
weighted dice similarity;
teks mining.

Abstract

The high similarity rate among undergraduate thesis titles has become a critical issue in maintaining the originality of academic work within higher education institutions. This study aims to develop an automated system for detecting title similarity by combining the Term Frequency–Inverse Document Frequency (TF-IDF) algorithm with the Weighted Dice Similarity method. TF-IDF is used to assign weights to important words in the titles, while Weighted Dice Similarity measures the degree of similarity between titles based on the distribution and weights of these words. The study utilizes a dataset of 200 manually annotated thesis titles as ground truth. The analysis process includes preprocessing, word weighting, and similarity computation between titles. Experimental results show that the system achieves an accuracy of 94%, a precision of 66.67%, a recall of 81.3%, and an average Weighted Dice similarity score of 0.62. Although the precision score is relatively moderate, the combination of both methods is considered effective, as it captures both lexical structure and semantic similarity—capabilities that are not fully achieved when using a single method alone.

***Penulis korespondensi:**
Santi Purwaningrum
E-mail: santi.purwaningrum@pnc.ac.id

1. Pendahuluan

Tugas akhir merupakan salah satu syarat kelulusan bagi mahasiswa di perguruan tinggi yang mencerminkan kemampuan mahasiswa dalam menerapkan ilmu pengetahuan dan keterampilan yang telah diperoleh selama masa studi. Tugas akhir mahasiswa berfungsi sebagai jembatan antara pendidikan formal dan dunia profesional, mempersiapkan mereka untuk menjadi profesional yang kompeten dan adaptif terhadap perkembangan teknologi dan informasi yang terus berubah [1]. Tugas akhir tidak hanya menjadi media evaluasi kompetensi, tetapi juga berfungsi sebagai kontribusi mahasiswa dalam pengembangan keilmuan melalui penelitian. Melalui tugas akhir, mahasiswa tidak hanya belajar untuk menyelesaikan proyek penelitian, tetapi juga memperoleh keterampilan penting lainnya, seperti kemampuan berpikir kritis, analisis, dan komunikasi yang efektif. Salah satu komponen yang sangat penting dalam penyusunan tugas akhir adalah pemilihan judul, karena judul menjadi representasi utama dari arah dan cakupan penelitian.

Judul tugas akhir harus bersifat orisinal, relevan dengan bidang studi, serta memiliki nilai kebaruan. Namun, dalam praktiknya, banyak mahasiswa mengalami kesulitan dalam menentukan judul yang benar-benar baru dan belum pernah digunakan sebelumnya. Akibatnya, sering ditemukan judul-judul tugas akhir yang memiliki kemiripan tinggi satu sama lain, baik dalam bentuk struktur kalimat, topik, maupun kata kunci. Fenomena ini diperparah dengan keterbatasan mahasiswa dalam melakukan pencarian literatur atau referensi judul yang komprehensif.

Tingginya tingkat kemiripan judul tugas akhir dapat menimbulkan berbagai permasalahan. Pertama, hal ini menimbulkan kekhawatiran terkait indikasi plagiarisme akademik, meskipun tidak selalu disengaja. Kedua, kemiripan judul berpotensi menimbulkan redundansi penelitian yang merugikan dalam konteks pengembangan ilmu pengetahuan. Ketiga, bagi dosen pembimbing dan pihak program studi, proses pengecekan kesamaan judul yang masih dilakukan secara manual menjadi kurang efisien [2].

Penelitian tentang deteksi kesamaan judul tugas akhir telah berkembang pesat dengan menggunakan berbagai pendekatan dan algoritma untuk mengatasi permasalahan plagiarisme. Penelitian mengenai deteksi judul tugas akhir banyak difokuskan pada metode pembobotan dan pengukuran kesamaan kata untuk meningkatkan efektivitas dalam mencegah plagiarisme dan membantu dalam pengelolaan data akademik. Salah satu studi yang relevan adalah penelitian oleh [3] yang mengembangkan sistem deteksi plagiarisme menggunakan *Natural Language Processing* (NLP) dengan algoritma *Jaro-Winkler* dan *Term Frequency-Inverse Document Frequency* (TF-IDF). Penelitian ini menunjukkan bahwa kombinasi algoritma tersebut dapat secara efektif mendeteksi kesamaan dalam penulisan tugas akhir mahasiswa, sehingga membantu menjaga integritas akademik dan mencegah plagiarisme.

Selain itu, penelitian oleh [4] menganalisis penerapan metode *Winnowing* untuk mendeteksi kesamaan judul tugas akhir. Hasil dari penelitian ini menunjukkan bahwa nilai parameter k yang digunakan dalam algoritma memiliki pengaruh signifikan terhadap hasil kesamaan yang terdeteksi, dimana variabel *k* lebih berpengaruh dibandingkan variabel lain dalam menentukan tingkat kesamaan antara judul-judul. Melalui penelitian-penelitian ini, jelas bahwa metode pembobotan dan pengukuran kesamaan seperti algoritma yang relevan, memainkan peran penting dalam meminimalkan plagiarisme dan memfasilitasi pengelolaan tugas akhir yang lebih efisien.

Penggunaan *Weighted Dice Similarity* dalam konteks tugas akhir mahasiswa telah dieksplorasi yang menunjukkan bahwa pendekatan ini meningkatkan efektivitas dalam pengambilan keputusan berdasarkan kriteria yang berbeda [5]. Hasil penelitian menunjukkan bahwa dengan menerapkan metode ini, sistem tidak hanya dapat mendeteksi kesamaan secara lebih akurat, tetapi juga dapat melakukan klasifikasi judul tugas akhir dengan mempertimbangkan relevansi setiap istilah yang terlibat, yang sangat penting dalam mencegah plagiarisme.

Penelitian penerapan algoritma text mining dan tf-idf untuk menganalisis abstrak skripsi mengembangkan sistem aplikasi yang memanfaatkan metode TF-IDF untuk mengelompokkan topik skripsi berdasarkan abstrak mahasiswa di Perpustakaan Universitas Dehasen Bengkulu. Penelitian ini bertujuan membantu petugas perpustakaan dalam mengelompokkan skripsi secara otomatis ke dalam beberapa kategori topik seperti Sistem Pakar, Data Mining, Sistem Pendukung Keputusan, dan Jaringan, sehingga memudahkan pencarian dan pengelolaan dokumen. Proses pengolahan teks melalui tahapan tokenizing, filtering, dan stemming menghasilkan bobot kata menggunakan TF-IDF yang kemudian digunakan untuk menentukan tingkat kemiripan dan pengelompokan abstrak. Hasil pengujian dengan data training 25 abstrak dan data testing 3 abstrak menunjukkan bahwa sistem mampu mengelompokkan skripsi dengan akurat, misalnya 2 skripsi masuk ke topik Data Mining (66,77%) dan 1 skripsi ke Sistem Pendukung Keputusan (33,33%). Kesimpulannya, aplikasi ini berjalan sesuai harapan dan efektif membantu pengelompokan topik skripsi secara otomatis, sehingga memberikan kemudahan bagi petugas perpustakaan dan pengguna dalam menemukan skripsi yang relevan [6].

Pada penelitian ini digunakan pendekatan teknik text mining dengan mengukur kemiripan antar judul menggunakan algoritma tertentu. Algoritma TF-IDF banyak digunakan dalam bidang pencarian informasi untuk menilai bobot pentingnya suatu kata dalam sebuah dokumen dibandingkan dengan dokumen lain. TF-IDF menghitung skor berdasarkan frekuensi kemunculan kata dalam dokumen dan seberapa unik kata tersebut di seluruh korpus dokumen, sehingga membantu dalam mengekstraksi fitur penting dari teks secara efektif dan efisien [7]. Selain itu, algoritma *Weighted Dice Similarity* digunakan untuk mengukur kesamaan antara dua set data dengan memperhitungkan bobot masing-masing elemen, yang relevan dalam analisis teks dan klasifikasi. *Weighted Dice* mengadaptasi rumus klasik *Dice similarity* dengan memasukkan bobot relevansi elemen yang dapat diperoleh dari TF-IDF.

Penelitian ini menggunakan pendekatan melalui teknik *text mining* dengan mengukur kemiripan antar judul

menggunakan algoritma tertentu. Algoritma TF-IDF banyak digunakan dalam bidang pencarian informasi untuk menilai bobot pentingnya suatu kata dalam sebuah dokumen dibandingkan dengan dokumen lain. TF-IDF menghitung skor berdasarkan frekuensi kemunculan kata dalam dokumen dan seberapa unik kata tersebut di seluruh korpus dokumen, sehingga membantu dalam mengekstraksi fitur penting dari teks [8]. Sementara itu, algoritma *Weighted Dice Similarity* adalah pendekatan yang digunakan untuk mengukur kesamaan antara dua set data, dengan memperhitungkan bobot dari masing-masing elemen dalam perhitungan. Ini sangat relevan dalam konteks pengolahan data, terutama dalam aplikasi yang memerlukan analisis teks dan klasifikasi. *Weighted Dice Similarity* mengadaptasi rumus klasik *Dice similarity*, yang dihitung dengan memperhatikan seberapa banyak elemen yang sama di antara dua set. Namun, dalam metode ini, setiap elemen diberi bobot berdasarkan relevansinya, yang dapat diperoleh melalui teknik seperti TF-IDF.

Penelitian ini bertujuan untuk mengisi *research gap* dengan mengembangkan pendekatan kombinasi yang mampu memberikan hasil evaluasi kemiripan judul yang lebih akurat dan kontekstual. Penelitian dalam penggunaan TF-IDF dan algoritma lain seperti *Weighted Dice Similarity* dapat dilihat dari pemahaman dan penerapan kombinasi keduanya. Sementara penelitian sebelumnya cenderung fokus pada penggunaan algoritma TF-IDF dalam pengukuran kemiripan antar dokumen atau judul tugas akhir mahasiswa, banyak yang belum mengeksplorasi potensi penggabungan algoritma tersebut untuk memperbaiki akurasi dan efisiensi dalam analisis data teks.

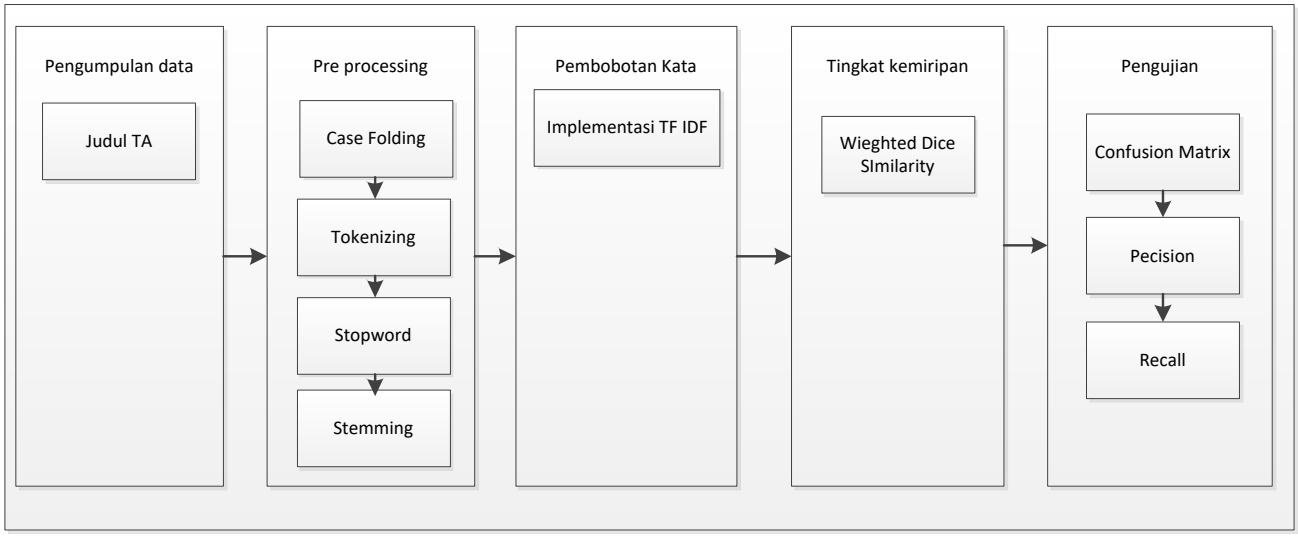
Kelebihan metode ini terletak pada kemampuannya menggabungkan pembobotan kata melalui TF-IDF dengan pengukuran tumpang tindih melalui *Weighted Dice Similarity*, yang memperhatikan bobot kontribusi dari setiap kata. Dibandingkan dengan pendekatan-pendekatan klasik sebelumnya, metode ini lebih kontekstual dan adaptif terhadap variasi redaksional. Selain itu, penelitian ini juga menghasilkan dataset *ground truth* yang dapat dimanfaatkan untuk validasi dan pengujian lanjutan sistem serupa. Dengan menggabungkan metode TF-IDF sebagai teknik ekstraksi fitur dan *Weighted Dice Similarity* sebagai pengukur kedekatan antar judul, diharapkan dapat diperoleh sistem yang mampu memberikan hasil deteksi kemiripan yang lebih optimal, terutama pada teks pendek seperti judul tugas akhir. Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi kombinasi kedua algoritma tersebut dalam konteks pendidikan tinggi, khususnya dalam membantu proses validasi judul tugas akhir mahasiswa secara otomatis.

2. Metode

Bagian ini menguraikan metodologi penelitian yang diterapkan untuk mengukur kemiripan judul tugas akhir menggunakan kombinasi algoritma TF-IDF dan *Weighted Dice Similarity*. Secara garis besar, metodologi ini terstruktur dalam enam tahapan utama: (1) Perancangan Penelitian, yang meliputi penentuan tujuan penelitian, identifikasi variabel, dan perumusan hipotesis; (2) Pengumpulan Data, yaitu proses pengumpulan dataset judul tugas akhir yang akan dianalisis; (3) Perancangan Data, yang melibatkan persiapan dan format data yang sesuai untuk diproses oleh algoritma; (4) Implementasi Metode TF-IDF, digunakan untuk menghitung bobot setiap kata dalam judul tugas akhir; (5) Implementasi *Metode Weighted Dice Similarity*, yang digunakan untuk mengukur kemiripan antar judul berdasarkan bobot kata yang dihasilkan oleh TF-IDF; dan (6) Metode Pengujian, yang mencakup evaluasi kinerja model menggunakan metrik yang relevan untuk mengukur akurasi dan efektivitas kombinasi algoritma TF-IDF dan *Weighted Dice Similarity* .

2.1 Perancangan Penelitian

Rancangan kegiatan penelitian terdiri dari beberapa tahapan utama, yaitu pengumpulan data, *preprocessing*, pembobotan kata, mengukir tingkat kemiripan, serta pengujian dan evaluasi hasil tingkat kemiripan antar judul. Tahapan penelitian di jelaskan pada gambar 1.



Gambar 1. Metode Penelitian

Berdasarkan kerangka pikir pada gambar 1, alur penelitian ini dimulai dengan pengumpulan data. Data mentah ini kemudian diproses melalui serangkaian tahapan *preprocessing* yang meliputi: *Case Folding* untuk mengubah semua teks menjadi huruf kecil, *Tokenizing* untuk memecah teks menjadi unit-unit kata atau token, penghapusan *Stopword* yaitu untuk menghapus kata-kata umum yang tidak memiliki nilai informatif, dan *Stemming* untuk mengubah kata ke bentuk dasarnya. Setelah itu, dilakukan pembobotan kata dengan Implementasi TF-IDF untuk menghitung bobot setiap kata dalam judul. Selanjutnya, tingkat kemiripan antar judul diukur menggunakan algoritma *Weighted Dice Similarity*. Terakhir, dilakukan pengujian dengan menghitung *Confusion Matrix, Precision*, dan *Recall* untuk mengevaluasi kinerja model dalam mengukur kemiripan judul.

Operasional variabel dalam penelitian ini dibagi menjadi dua: (1) Variabel bebas yaitu algoritma yang digunakan (TF-IDF dan Weighted Dice Similarity), dan (2) Variabel terikat yaitu skor kemiripan judul yang dihasilkan oleh sistem. Skor ini berkisar antara 0 sampai 1, di mana nilai yang lebih tinggi menunjukkan tingkat kemiripan yang lebih besar.

2.2 Pengumpulan data

Data berupa kumpulan judul tugas akhir mahasiswa yang diperoleh dari repositori kampus dalam rentang lima tahun terakhir. Tahap pengumpulan data merupakan langkah awal yang krusial dalam penelitian ini. Data yang digunakan adalah kumpulan judul tugas akhir dari repositori kampus dalam rentang lima tahun terakhir. Pemilihan judul TA sebagai unit analisis didasarkan pada pertimbangan bahwa judul TA merepresentasikan inti dari topik penelitian yang dilakukan oleh mahasiswa, sehingga kemiripan judul dapat mengindikasikan adanya kesamaan topik atau tema penelitian. Proses pengumpulan data dilakukan dengan ekstraksi dari database. Dataset yang terkumpul kemudian disimpan dalam format untuk memudahkan proses preprocessing dan analisis selanjutnya.

2.3 Perancangan Data

Perancangan data merupakan tahap penting dalam memastikan bahwa data yang digunakan dalam penelitian siap untuk diproses secara optimal oleh algoritma yang diterapkan. Data yang digunakan dalam penelitian ini berupa kumpulan judul tugas akhir yang diperoleh dari repositori akademik perguruan tinggi. Sebelum data dianalisis, dilakukan serangkaian proses *preprocessing* untuk membersihkan dan menstandarkan format teks, guna meningkatkan akurasi analisis kemiripan.

Setelah data judul tugas akhir terkumpul masuk pada tahap pertama yaitu *text preprocessing*. Metode *preprocessing* dalam *text mining* sangat penting untuk memastikan data judul tugas akhir mahasiswa dapat diproses dengan optimal sebelum analisis lebih lanjut. Kegiatan *preprocessing* mencakup beberapa langkah yang berfokus pada membersihkan dan menyiapkan data untuk analisis. Secara umum, tahapan-tahapan tersebut meliputi *case folding*, penghapusan simbol dan karakter non-alfabet, penghilangan *stopword*, tokenisasi, dan *stemming* [9]. *Case folding* adalah proses mengubah semua karakter menjadi huruf kecil untuk menghilangkan perbedaan akibat penggunaan huruf besar atau kecil, sehingga menjaga konsistensi dalam analisis teks. *Stopword removal* yaitu penghapusan simbol dan karakter non-alfabet dilakukan untuk memastikan bahwa teks yang diproses tidak mengandung elemen yang dapat mengganggu analisis, seperti tanda baca atau angka yang tidak relevan [10]. Penghilangan *stopword* juga merupakan langkah penting karena *stopword* adalah kata-kata umum yang tidak memiliki kontribusi makna signifikan dalam konteks analisis [11]. Teknik ini dapat mempengaruhi efisiensi dan efektivitas hasil analisis, dengan memperkecil ukuran data yang perlu diproses, sehingga mengurangi waktu komputasi [12]. Tokenisasi adalah proses memecah teks menjadi unit individual, atau "token", seperti kata atau frasa, yang selanjutnya digunakan dalam analisis. Stemming adalah proses mengembalikan kata ke bentuk dasarnya untuk menyatukan variasi kata yang berbeda yang memiliki makna serupa; misalnya, "berjalan", "berjalanlah", dan "berjalan-jalan" akan distemming menjadi "jalan" [13] [14].

Data yang telah melalui tahap *preprocessing* kemudian direpresentasikan dalam bentuk matriks vektor, dengan bobot yang dihitung melalui algoritma TF-IDF. Matriks ini selanjutnya digunakan sebagai input dalam perhitungan kemiripan menggunakan metode *Weighted Dice Similarity*. Dengan perancangan data yang terstruktur ini, penelitian dapat memastikan bahwa informasi yang relevan dalam setiap judul terwakili secara optimal dalam proses analisis dan evaluasi kesamaan teks

2.4 Metode TF-IDF

Setelah teks dibersihkan, data kemudian diubah menjadi representasi numerik menggunakan algoritma TF-IDF. Metode pembobotan kata dengan algoritma TF-IDF adalah teknik yang banyak digunakan dalam pengolahan teks dan analisis informasi. Secara mendasar, TF-IDF bekerja berdasarkan dua komponen utama: frekuensi term dan frekuensi dokumen, pada persamaan (1).

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

Dimana :

- TF (*Term Frequency*) menghitung seberapa sering suatu kata muncul di dalam dokumen. Biasanya, ini dinyatakan dalam bentuk persamaan (2):

$$TF(t,d) = \frac{jumlah\ kemunculan\ term\ t\ dalam\ dokumen\ d}{Jumlah\ seluruh\ term\ dalam\ dokumen\ d} \quad (2)$$

- IDF (Inverse Document Frequency) mengukur pentingnya suatu kata di seluruh koleksi dokumen. Rumusnya adalah pada persamaan (3).

$$IDF(t) = \log \frac{Total\ jumlah\ dokumen}{Jumlah\ dokumen\ yang\ mengandung\ term\ t} \quad (3)$$

Frekuensi term (TF) mengukur seberapa sering sebuah kata muncul dalam sebuah dokumen, memberikan bobot yang lebih tinggi pada kata-kata yang sering muncul dalam konteks dokumen tersebut. Di sisi lain, frekuensi dokumen (IDF) mengevaluasi seberapa umum atau jarangnya kata tersebut dalam kumpulan dokumen. Dengan membagi jumlah total dokumen dengan jumlah dokumen yang mengandung kata tertentu dan mengambil logaritma dari hasilnya, IDF memberikan penekanan pada kata-kata yang lebih spesifik dan relevan dalam konteks keseluruhan [15] [16].

2.5 Weighted Dice Similarity

Metode *Weighted Dice Similarity* adalah salah satu pendekatan yang digunakan untuk mengukur kesamaan antara dua set data dengan memberikan bobot pada setiap elemen dalam perhitungan. Pendekatan ini memperluas konsep asli dari Dice similarity, yang mendasari pengukuran kesamaan dua set dengan rumus dasar yang mengacu pada jumlah elemen yang sama, dengan menambahkan bobot yang mencerminkan relevansi setiap elemen dalam konteks yang lebih luas [17] [18]. Dengan menggunakan bobot, metode ini dapat menyoroti elemen-elemen yang lebih penting, mengatasi kelemahan metode *Dice* biasa yang dapat menempatkan semua elemen pada tingkat kepentingan yang sama, meskipun beberapa elemen mungkin memiliki dampak yang lebih besar pada keseluruhan kesamaan, pada persamaan (4).

$$Dice_{Weighted} = \frac{2 \times \sum_{\omega \in A \cap B} \min\left(tfidf_A(\omega), \sum_{\omega \in A \cap B} \min\left(tfidf_A(\omega)\right)\right)}{\sum_{\omega \in A} tfidf_A(\omega) + \sum_{\omega \in B} tfidf_B(\omega)} \quad (4)$$

- *A,B*: Mewakili dua dokumen yang akan dibandingkan.
- ω: Menunjukkan sebuah token atau kata dalam dokumen.
- $A \cap B$: Irisan dari dua dokumen, yaitu himpunan kata yang muncul di kedua dokumen.
- $tfidf_A(\omega)$: Nilai TF-IDF dari kata ω dalam dokumen A. Nilai ini mencerminkan seberapa penting kata tersebut dalam dokumen A.
- $tfidf_B(\omega)$: Nilai TF-IDF dari kata ω dalam dokumen B.
- $\min\left(tfidf_A(\omega), tfidf_B(\omega)\right)$: Memilih nilai terkecil dari TF-IDF kata ω pada dokumen A dan B, sebagai bentuk konservatif dari kontribusi kemiripan (karena kata harus penting di kedua dokumen untuk memberikan kontribusi besar).
- $\sum_{\omega \in A \cap B}$: Menjumlahkan nilai minimum TF-IDF dari setiap kata ω yang ada di kedua dokumen.
- $\sum_{\omega \in A} tfidf_A(\omega)$ dan $\sum_{\omega \in B} tfidf_B(\omega)$: Menjumlahkan seluruh bobot TF-IDF dari kata-kata dalam masing-masing dokumen.

Rumus tersebut menghitung kemiripan dengan memberi penekanan lebih pada kata-kata penting (berdasarkan TF-IDF), dan membandingkan seberapa banyak dan seberapa penting kata-kata yang sama dalam kedua dokumen, dibandingkan dengan total bobot kata di masing-masing dokumen [19].

2.6 Pengujian

Dalam menilai kinerja sistem yang telah dibangun, evaluasi dilakukan dengan menggunakan metode *confusion matrix*, yang merupakan alat penting dalam pengukuran efektivitas sistem klasifikasi. *Confusion matrix* memberikan gambaran menyeluruh terkait hasil klasifikasi, yang terdiri dari empat elemen utama: *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), dan *False Negatives* (FN) [20]. Dari elemen-elemen ini, dua metrik utama yang sering digunakan adalah *precision* dan *recall*. *Precision* mengukur proporsi prediksi positif yang benar dibandingkan dengan total prediksi positif. Hal ini penting karena *precision* menggambarkan seberapa akurat model dalam memberikan prediksi positif. Di sisi lain, *recall* mengukur proporsi prediksi positif yang benar dibandingkan dengan total sebenarnya dari kelas positif. Rumus *Precision* dan *Recall* dinyatakan pada persamaan (5) dan persamaan (6).

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

TP merupakan Jumlah prediksi positif yang benar, sedangkan FP adalah Jumlah prediksi positif yang salah kenyataannya "negatif". FN Jumlah kasus positif yang gagal diprediksi sebagai positif (sistem mengira negatif padahal sebenarnya positif. Dengan tahapan analisis ini, sistem mampu memberikan hasil deteksi kemiripan judul yang efektif dan dapat digunakan sebagai alat bantu dalam menjaga orisinalitas karya ilmiah mahasiswa.

3. Hasil dan Pembahasan

3.1 Hasil Penelitian

Hasil dari penelitian ini disajikan secara kuantitatif berdasarkan tahapan kegiatan yang telah dilakukan, mulai dari analisis data judul, perancangan sistem, hingga proses pengujian performa algoritma. Pada tahap awal, dilakukan analisis terhadap sekumpulan data judul tugas akhir mahasiswa yang diambil dari repositori kampus dalam format digital.

a) Preprocessing

Dari total 231 judul yang dikumpulkan, dilakukan proses *preprocessing* melalui teknik *tokenisasi*, *case folding*, *stopword removal*, dan *stemming* untuk menghasilkan representasi teks yang bersih dan siap dianalisis lebih lanjut. Tujuan utama dari *preprocessing* adalah untuk mengubah teks mentah menjadi bentuk representasi yang lebih bersih, terstruktur, dan mudah diolah secara komputasional.

Masing-masing tahapan *preprocessing* memiliki peranan penting dalam menyederhanakan dan menormalkan teks tanpa menghilangkan makna utamanya. Tahap pertama adalah tokenisasi, yaitu proses memecah teks menjadi satuan-satuan kecil yang disebut *token*, biasanya berupa

kata. Tokenisasi membantu mesin untuk memahami teks dalam unit-unit dasar yang bisa dianalisis lebih lanjut.
Contoh :

APLIKASI PEMBELAJARAN METAMORFOSIS SERANGGA MENGGUNAKAN AUGMENTED REALITY BERBASIS ANDROID (STUDI KASUS SDN JERUKLEGI WETAN 01)

Setelah proses tokenisasi :

"APLIKASI", "PEMBELAJARAN", "METAMORFOSIS", "SERANGGA", "MENGGUNAKAN", "AUGMENTED", "REALITY", "BERBASIS", "ANDROID", "STUDI", "KASUS", "SDN", "JERUKLEGI", "WETAN", "01"

Tahap selanjutnya adalah *case folding*, yaitu proses mengubah semua huruf menjadi huruf kecil. Hal ini dilakukan untuk menyamakan bentuk kata yang sama namun berbeda kapitalisasi. Hasil *case folding* menjadi :

"aplikasi", "pembelajaran", "metamorfosis", "serangga", "menggunakan", "augmented", "reality", "berbasis", "android", "studi", "kasus", "sdn", "jeruklegi", "wetan", "01"

Tahap ketiga adalah *stopword removal*, yaitu menghapus kata-kata umum yang tidak membawa makna penting dalam analisis, seperti "yang", "dan", "dengan", "untuk", dll. Pada judul contoh, kata-kata seperti “menggunakan”, “studi”, dan “kasus” bisa saja dianggap sebagai *stopword* tergantung pada kamus *stopword* yang digunakan. Hasil *stopword* removal menjadi:

"aplikasi", "pembelajaran", "metamorfosis", "serangga", "augmented", "reality", "berbasis", "android", "sdn", "jeruklegi", "wetan", "01"

Tahap terakhir adalah *stemming*, yaitu proses mengembalikan kata ke bentuk dasarnya. Proses ini penting untuk menyamakan variasi kata dengan akar kata yang sama. Misalnya, “pembelajaran” diubah menjadi “ajar”, “berbasis” menjadi “basis”. Dengan demikian, hasil stemming menjadi:

"aplikasi", "ajar", "metamorfosis", "serangga", "augmented", "reality", "basis", "android", "sdn", "jeruklegi", "wetan", "01"

Setelah melalui keempat tahapan *preprocessing* tersebut, teks telah direduksi menjadi representasi yang lebih sederhana namun tetap kaya informasi. Representasi ini memungkinkan proses analisis lanjutan seperti klasifikasi dokumen, pencocokan topik, atau pengelompokan konten menjadi lebih efisien dan akurat.

b) Pembobotan kata

Pada tahap ini, dilakukan proses pembobotan terhadap setiap kata yang terdapat dalam kumpulan judul tugas akhir menggunakan algoritma TF-IDF Pembobotan ini bertujuan untuk mengukur tingkat kepentingan suatu kata dalam sebuah dokumen relatif terhadap keseluruhan korpus. Kata-kata yang sering muncul dalam satu dokumen tetapi jarang muncul di dokumen lain akan memiliki bobot TF-IDF yang lebih tinggi, karena dianggap lebih representatif terhadap isi dokumen tersebut. Sebaliknya, kata-kata umum yang muncul di banyak dokumen akan mendapatkan bobot yang lebih rendah karena memiliki tingkat diskriminasi yang kecil. Hasil dari pembobotan ini menjadi dasar penting dalam proses perhitungan kemiripan teks pada tahap selanjutnya, khususnya saat diterapkannya algoritma *Weighted Dice Similarity*, yang mempertimbangkan bobot TF-IDF sebagai nilai kontribusi antar kata.

Tahapa wal TF-TDF adalah mengukur seberapa sering suatu kata muncul dalam satu dokumen yang biasa disebut TF dengan cara Jumlah kemunculan kata dibagi dengan jumlah total kata. Pada contoh kata setelah di *processing* diatas terdapat 12 kata berasal dari satu dokumen, dan tiap kata muncul 1 kali, maka TF tiap kata:

$$TF = \frac{1}{12} = 0{,}083$$

Kemudian masuk pada proses IDF yang berfungsi mengukur seberapa penting suatu kata secara keseluruhan. Kata yang sering muncul di banyak dokumen dianggap kurang informatif, sehingga IDF-nya kecil, IDF dapat dihitung dengan cara log dari jumlah total dokumen dibagi dengan jumlah dokumen dalam korpus yang mengandung kata tertentu.

$$IDF\ \text{"aplikasi"} = \log\frac{415}{325} = 0.106165$$
$$IDF\ \text{"jeruklegi"} = \log\frac{415}{11} = 37.78$$

Setelah menghitung TF dan IDF, nilai TF-IDF dihitung dengan mengalikan keduanya:

TFI-DF "aplikasi"=0,083 x 0.106165 = 0.008812

TFI-DF "aplikasi"=0,083 x 37.78 = 3.131363636

Dari contoh ini terlihat bahwa kata seperti "jeruklegi", yang jarang muncul di seluruh dokumen, memiliki nilai TF-IDF lebih tinggi, sehingga dianggap lebih informatif atau khas untuk dokumen tersebut. Sedangkan kata seperti "aplikasi" yang umum, meskipun muncul, kontribusinya terhadap representasi dokumen akan lebih kecil.

Hasil dari pembobotan ini menjadi dasar penting dalam proses perhitungan kemiripan teks pada tahap selanjutnya, khususnya saat diterapkannya algoritma *Weighted Dice Similarity*, yang mempertimbangkan bobot TF-IDF sebagai nilai kontribusi antar kata.

c) Tingkat kemiripan

Salah satu pendekatan untuk mengukur kemiripan semantik antar judul adalah dengan menggunakan metode *Weighted Dice Similarity*, yang menggabungkan representasi kata melalui teknik vektorisasi seperti TF-IDF dengan pengukuran overlap antar kata-kata kunci. Tujuan dari perhitungan ini adalah untuk mengetahui sejauh mana

kedua judul tersebut memiliki kemiripan konten secara semantik berdasarkan keterhubungan kata-kata penting yang dimiliki masing-masing. Contoh:

Judul A : “aplikasi", "ajar", "metamorfosis", "serangga", "augmented", "reality", "basis", "android", "sdn", "jeruklegi", "wetan", "01"

Judul B : “aplikasi”, “ajar”, “planet”, “guna”, “augmented”, “reality, “basis”, “smartphone”, “android”

Dengan perhitungan menggunakan rumus *Weighted Dice Similarity* dari dua judul diatas adalah sebegai berikut: Intersection (∑ min) adalah Jumlah minimum dari nilai TF-IDF untuk kata-kata yang muncul di kedua judul.
= min (aplikasi) + min (ajar) + min (augmented) + min (reality) + min (basis) + min (android) = 6 × 0.2191 = 1.3146

Total IDF adalah $\sum TFIDF_A$ =2.6308, $\sum TFIDF_B$ =0.4913, Total = 5.3526

Sehingga *Weighted Dice Similarity* = (2x1.3146) / 5.3526 = 0.4913

Hasil perhitungan *Weighted Dice Similarity* berbasis TF-IDF menunjukkan bahwa kemiripan antara kedua judul adalah 0.4913 atau 49.13%. Ini mengindikasikan bahwa meskipun judul-judul ini memiliki fokus topik berbeda (“metamorfosis serangga” vs. “planet”), terdapat kemiripan dalam penggunaan istilah teknologi dan struktur kalimat. Metode ini efektif dalam mengukur kemiripan semantik berbasis bobot kata nyata, dan sangat cocok untuk aplikasi seperti pencarian dokumen mirip, penyaringan duplikasi, atau klasifikasi topik, seperti pada Tabel 1.

Tabel 1. Vektorisasi menggunakan TF-IDF dan similarity menggunakan Weighted

No	Judul A	Judul B	Weighted Dice
1	aplikasi pembelajaran metamorfosis serangga...	aplikasi pembelajaran planet...	0.42
2	aplikasi pembelajaran metamorfosis serangga...	media pembelajaran interaktif gerakan dan bacaan wudhu...	0.45
3	aplikasi pembelajaran planet...	media pembelajaran interaktif gerakan dan bacaan wudhu...	0.37
4	penerapan teknologi augmented reality untuk senam ibu hamil...	aplikasi pembelajaran planet...	0.16
5	perancangan pembelajaran gerakan sholat...	media pembelajaran interaktif gerakan dan bacaan wudhu...	0.39

Judul 1 & 3 memiliki nilai *similarity* 0.45, menandakan banyak tumpang tindih kata kunci seperti “pembelajaran”, “menggunakan”, dan “augmented reality”. Judul 2 & 3 memiliki similarity 0.37, menunjukkan kemiripan tema meski topiknya berbeda (planet vs wudhu). Judul 1 & 4 hanya 0.19, artinya meskipun sama-sama menggunakan "augmented reality", konten tematik sangat berbeda (serangga vs senam ibu hamil). Nilai *similarity* berada di rentang 0.16–0.45, menunjukkan adanya kesamaan kata namun juga keberagaman isi, seperti pada Tabel 2.

Tabel 2. Lima Pasangan Non-Identik dengan Similarity Tertinggi

No.	Judul A	Judul B	Similaritas
1	sistem informasi pemesanan dan penjualan berbasis android	sistem informasi pemesanan dan penjualan berbasis android	0.92
2	sistem informasi penjualan hasil laut berbasis android	sistem informasi penjualan hasil laut berbasis android	0.89
3	aplikasi pengelolaan data aset sekolah berbasis android	aplikasi pepngelolaan data aset sekolah berbasis android	0.85
4	sistem informasi pengelolaan anggota kwartir cabang berbasis web	sistem informasi pengelolan anggota pramuka kwartir cabang berbasis web	0.80
5	media pembelajaran interaktif gerakan dan bacaan wudhu menggunakan augmented reality	media pembelajaran interaktif gerakan dan bacaan wudhu menggunakan augmented reality berbasis android	0.73

Perbedaan kecil dalam ejaan, seperti pengelolaan vs pepngelolaan, tetap berdampak pada skor, tetapi *similarity* tetap tinggi karena kata lainnya identik. Pasangan ke-5 memperlihatkan bahwa penambahan frasa kecil seperti "berbasis android" tidak mengubah makna secara signifikan, tapi memengaruhi bobot TF-IDF, seperti pada Tabel 3.

Tabel 3. Pasangan Judul dengan Similarity Terendah (0.00)

No.	Judul A	Judul B
1	rancang bangun aplikasi pencatat stok barang di toko bahan bangunan berbasis web	sistem informasi penjualan pada pangkalan gas elpiji
2	rancang bangun aplikasi pencatat stok barang di toko bahan bangunan berbasis web	sistem penjualan pada rumah makan ayam penyet
3	aplikasi pembelajaran planet menggunakan augmented reality berbasis smartphone android	sistem informasi penjualan pada pangkalan gas elpiji
4	rancang bangun aplikasi	sistem pendukung

No.	Judul A	Judul B
	pencatat stok barang di toko bahan bangunan berbasis web	keputusan penentuan beasiswa siswa baru
5	media pembelajaran interaktif gerakan dan bacaan wudhu menggunakan augmented reality	sistem informasi penilaian perkembangan anak di paud berbasis android

Judul-judul ini berasal dari domain yang sangat berbeda: pembelajaran vs penjualan, atau pendidikan agama vs sistem distribusi barang. Tidak ada istilah atau frasa penting yang tumpang tindih, sehingga similarity-nya benar-benar nol. Judul-judul seperti "augmented reality" atau "penilaian perkembangan anak" tidak relevan satu sama lain secara semantik maupun teknis.

Setelah dilakukan transformasi, dari setiap judul diambil lima token dengan nilai TF-IDF tertinggi untuk kemudian digunakan sebagai masukan dalam perhitungan *Dice Similarity*. *Dice Similarity* digunakan untuk menilai sejauh mana dua judul memiliki token yang sama. Sistem dirancang untuk menghasilkan skor kemiripan antar judul dalam rentang 0 sampai 1, di mana nilai di atas *threshold* 0.75 dikategorikan sebagai judul yang berpotensi mirip secara ide atau topik, seeprti pada Tabel 4.

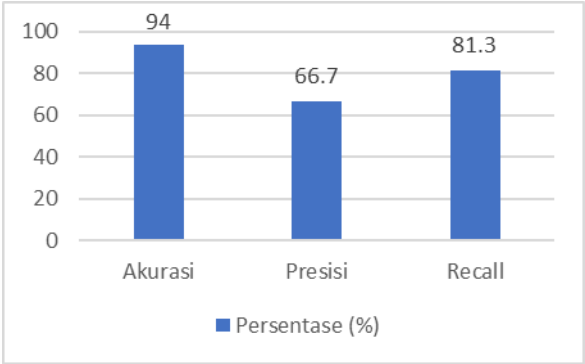
Tabel 4. Pasangan Judul dengan nilai kemiripan di atas ambang batas 0.75

No	Judul 1	Judul 2	Similarity
1	sistem informasi pengelolaan anggota kwartir cabang pramuka	sistem informasi pengelolan anggota pramuka kwartir cabang	0.8034
2	sistem informasi pemesanan dan penjualan berbasis android	sistem informasi pemesanan dan penjualan berbasis android	0.9229
3	sistem informasi penjualan hasil laut berbasis web	sistem informasi penjualan hasil laut berbasis web	0.8903

Dari hasil perhitungan, ditemukan bahwa beberapa judul memiliki nilai kemiripan di atas ambang batas 0.75, yang menunjukkan kemiripan yang signifikan baik dari segi struktur maupun konten. Misalnya, judul "sistem informasi pengelolaan anggota kwartir cabang pramuka" dan "sistem informasi pengelolan anggota pramuka kwartir cabang" memiliki nilai *similarity* sebesar 0.8031, yang menunjukkan bahwa keduanya hampir identik meskipun ada sedikit perbedaan penulisan kata. Demikian pula, judul "sistem informasi pemesanan dan penjualan berbasis android" muncul dua kali dengan kemiripan 0.9228, menunjukkan kemungkinan adanya duplikasi data. Satu lagi pasangan judul, "sistem informasi penjualan hasil laut berbasis web", juga muncul dua kali dengan nilai kemiripan 0.8893, menguatkan indikasi duplikasi atau pengulangan yang tidak disengaja.

d) Pengujian

Pada tahapan terakhir dari proses pembangunan sistem, dilakukan evaluasi kinerja menggunakan pendekatan *confusion matrix* untuk menilai sejauh mana sistem mampu mengidentifikasi kemiripan antar judul dengan akurat. Evaluasi ini mengelompokkan hasil prediksi sistem ke dalam empat kategori: *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN), dengan membandingkan hasil prediksi sistem terhadap label *ground truth* yang telah disusun berdasarkan kemiripan struktur awal judul dan nilai *similarity*. Sistem menganggap dua judul mirip apabila nilai *similarity* ≥ 0.75, dan *ground truth* menetapkan pasangan sebagai benar-benar mirip jika memenuhi syarat tersebut dan memiliki kesamaan struktur tiga kata pertama.



Gambar 2. Evaluasi kombinasi algoritma TF-IDF dan *Weighted Dice Similarity*

Berdasarkan perbandingan antara hasil prediksi dan data *ground truth* tersebut, diperoleh metrik evaluasi sebagai berikut:

- Akurasi sebesar 94% yang menunjukkan bahwa hampir seluruh pasangan judul diklasifikasikan dengan benar oleh sistem.
- Dalam analisis presisi sistem prediksi kemiripan judul, nilai presisi sebesar 0.6667 menunjukkan bahwa hanya sekitar 66,67% dari semua pasangan judul yang diprediksi benar-benar mirip memiliki kemiripan yang sesuai dengan kebenaran (*ground truth*). Meskipun angka ini tampak relatif tinggi, terdapat sejumlah *false positive*, yang berarti beberapa pasangan judul yang tidak benar-benar mirip tetap diprediksi sebagai mirip oleh sistem. Nilai presisi di bawah 65% dapat diartikan sebagai indikator bahwa sistem masih memiliki potensi untuk meningkatkan akurasi prediksinya. Ketidakakuratan ini dapat disebabkan oleh beberapa faktor, termasuk ketidakefektifan metode pengukuran kesamaan yang digunakan.
- *Recall* sebesar 81,3% mengindikasikan bahwa sistem cukup sensitif, namun masih terdapat sekitar 18,7% pasangan judul mirip yang tidak berhasil dikenali oleh sistem (*false negatives*). Hal ini bisa terjadi karena kemiripan teks secara semantik tidak selalu tercermin secara eksplisit dalam struktur kalimat atau urutan kata—terutama jika perbedaan redaksional, sinonim, atau struktur frasa tidak ditangani secara optimal oleh metode berbasis TF-IDF.

Secara keseluruhan, sistem menunjukkan performa yang sangat tinggi dalam mendeteksi kemiripan judul, sehingga sistem mampu mendeteksi judul yang mirip dengan cukup baik dan seimbang antara ketepatan dan kelengkapan. Temuan ini menunjukkan bahwa kombinasi TF-IDF dan Dice Similarity mampu mengurangi risiko duplikasi ide dan dapat digunakan sebagai alat bantu administratif untuk pengawasan orisinalitas topik tugas akhir.

3.2 Pembahasan

Pembahasan mengenai kombinasi algoritma TF-IDF dan *Weighted Dice Similarity* untuk menentukan tingkat kemiripan judul tugas akhir mahasiswa bertujuan untuk mengidentifikasi potensi kesamaan dalam penamaan tugas akhir di kalangan mahasiswa. Penelitian ini menyoroti kekurangan dan tantangan yang dihadapi dalam analisis teks, yang sering kali melibatkan pengukuran kesamaan yang tidak mendalam dalam konteks akademik. Hasil evaluasi menunjukkan bahwa sistem ini mampu mengenali pola kemiripan judul dengan akurasi tinggi, serta tingkat *recall* yang cukup baik, yaitu 81,3%, yang berarti mayoritas pasangan judul yang sebenarnya mirip berhasil diidentifikasi oleh sistem. Hal ini menunjukkan bahwa pendekatan yang digunakan efektif dalam konteks pemrosesan teks pendek seperti judul penelitian, meskipun masih terdapat beberapa pasangan mirip yang terlewatkan.

Hasil ini memperlihatkan adanya kebaruan pada aspek metode kombinasi, yaitu dengan memadukan TF-IDF dan *Weighted Dice Similarity* dalam proses evaluasi teks pendek. Dibandingkan dengan penelitian sebelumnya yang hanya mengandalkan cosine similarity atau metode *string matching* biasa, pendekatan ini memberikan fleksibilitas dan akurasi yang lebih baik, khususnya dalam mengidentifikasi teks yang berbeda secara struktural namun sama secara semantik. penelitian ini juga menghasilkan kontribusi praktis berupa dataset ground truth yang dapat digunakan untuk evaluasi sistem secara objektif. Penyusunan label referensi ini memungkinkan dilakukan evaluasi berbasis *confusion matrix* yang menghasilkan nilai akurasi sebesar 94%, presisi 66,67%, dan *recall* 81,3%. Nilai presisi yang masih perlu ditingkatkan mengindikasikan bahwa sistem sesekali mengklasifikasikan judul yang tidak terlalu mirip sebagai mirip.

Dengan demikian, pembahasan ini menunjukkan bahwa sistem yang dikembangkan tidak hanya memenuhi tujuan penelitian, tetapi juga menghadirkan kontribusi baru pada metode deteksi kemiripan judul berbasis teks pendek. Pendekatan gabungan yang digunakan terbukti dapat memberikan hasil yang lebih kontekstual dan akurat dibandingkan metode sebelumnya

4. Kesimpulan

Penelitian ini menyimpulkan bahwa kombinasi algoritma TF-IDF dan *Weighted Dice Similarity* terbukti efektif dalam mendeteksi tingkat kemiripan antar judul tugas akhir mahasiswa. Sistem yang dibangun mampu mengidentifikasi pasangan judul yang memiliki struktur atau makna yang serupa dengan cukup akurat, meskipun redaksional judul berbeda. Dari hasil pengujian judul tugas akhir yang telag dilakukan, sistem menunjukkan performa evaluasi yang baik, dengan nilai akurasi sebesar 94%, presisi sebesar 66,67%, dan recall sebesar 81,3%. Nilai recall yang tinggi menunjukkan bahwa sistem memiliki sensitivitas yang kuat dalam mengenali judul-judul yang memang mirip, meskipun masih terdapat kelemahan pada aspek presisi yang menunjukkan adanya prediksi mirip yang tidak sepenuhnya tepat.

Sebagai saran, sistem dapat dikembangkan lebih lanjut dengan mengintegrasikan pendekatan berbasis semantik seperti word embeddings (contoh: Word2Vec atau BERT) untuk meningkatkan presisi dan menangkap kesamaan makna yang tidak tergambar melalui kata-kata eksplisit. Sistem ini berpotensi besar untuk diterapkan secara luas dalam validasi judul tugas akhir di institusi pendidikan tinggi sebagai alat bantu administratif dalam menjaga orisinalitas dan mengurangi risiko duplikasi penelitian.

Ucapan Terimakasih

Pada bagian ini diisi ucapan terimakasih yang ditujukan kepada pihak yang telah memberikan bantuan dalam proses penelitian baik itu dari biaya penelitian, maupun dukungan lainnya.

Daftar Pustaka

[1] G. Rininda, I. H. Santi, and S. Kirom, "Penerapan SVM Dalam Analisis Sentimen Pada Edlink Menggunakan Pengujian Confusion Matrix," *Jati (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 5, pp. 3335–3342, 2024, doi: 10.36040/jati.v7i5.7420.

[2] M. F. Azima, A. N. Listanto, P. Studi, T. Informatika, F. Matching, and D. Plagiarisme, "Kombinasi Algoritma TF-IDF dan Fuzzy Matching untuk Deteksi Kemiripan Judul Skripsi," vol. 19, no. x, pp. 1–11, 2024.

[3] D. Darmanto, N. I. Pradasari, and E. Wahyudi, "Sistem Deteksi Plagiarisme Tugas Akhir Mahasiswa Berbasis Natural Language Processing Menggunakan Algoritma Jaro-Winkler Dan TF-IDF," *Smart Comp Jurnalnya Orang Pint. Komput.*, vol. 13, no. 1, 2024, doi: 10.30591/smartcomp.v13i1.6375.

[4] A. Y. Bramantya, T. Hasanuddin, and F. Umar, "Analisis Metode Winnowing Dalam Pendeteksian Plagiarisme Judul," *Bul. Sist. Inf. Dan Teknol. Islam*, vol. 3, no. 4, pp. 268–273, 2022, doi: 10.33096/busiti.v3i4.1469.

[5] Z. Ali and T. Mahmood, "Complex Neutrosophic Generalised Dice Similarity Measures and Their Application to Decision Making," *Caai Trans. Intell. Technol.*, vol. 5, no. 2, pp. 78–87, 2020, doi: 10.1049/trit.2019.0084.

[6] H. Sari, G. Leonarde Ginting, and T. Zebua, "Penerapan Algoritma Text Mining Dan Tf-Idf Untuk Pengelompokan Topik Skripsi," *Terap. Inform. Nusant.*, vol. 2, no. 7, pp. 414–432, 2021, [Online]. Available: https://ejurnal.seminar-id.com/index.php/tin

[7] R. Albin Pranata, N. Azmi Verdikha, U. Muhammdiyah Kalimantan Timur, and J. H. Ir Juanda, "Metode Pembobotan Tf-Idf Untuk Klasifikasi Teks Quick Count Pemilihan Wakil Presiden Indonesia 2024 Pada X Twitter Dengan Metode Svm," vol. 18, no. 2, p. 126, 2024, [Online]. Available:

https://doi.org/10.47111/JTIAvailableonlineathttps://e-journal.upr.ac.id/index.php/JTI

[8] J. A. Sari and B. A. Diana, "Dampak Transformasi Digitalisasi terhadap Perubahan Perilaku Masyarakat Pedesaan," *J. Pemerintah. dan Polit.*, vol. 9, no. 2, pp. 88–96, 2024, doi: 10.36982/jpg.v9i2.3896.

[9] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organ. Res. Methods*, vol. 25, no. 1, pp. 114–146, 2020, doi: 10.1177/1094428120971683.

[10] A. R. Purnajaya, V. Lieputra, V. Tayanto, and J. G. Salim, "Implementasi Text Mining Untuk Mengetahui Opini Masyarakat Tentang Climate Change," *J. Inf. Syst. Technol.*, vol. 3, no. 3, p. 36, 2022, doi: 10.37253/joint.v3i3.7337.

[11] F. Alshanik, A. Apon, A. Herzog, I. Safro, and J. Sybrandt, "Accelerating Text Mining Using Domain-Specific Stop Word Lists," pp. 2639–2648, 2020, doi: 10.1109/bigdata50022.2020.9378226.

[12] J. Gyarmati, P. Zentay, G. Gávay, and F. Hajdú, "Experimental Investigation and Statistical Analysis for Spallation Characteristics of Ballistic Penetration," *Acad. Appl. Res. Mil. Public Manag. Sci.*, vol. 18, no. 2, pp. 39–56, 2019, doi: 10.32565/aarms.2019.2.4.

[13] S. S. J, "A Novel Support Vector Machine Based Improved Aquila Optimizer-Based Text Mining Mechanism for the Healthcare Applications," vol. 20, no. 5s, pp. 2909–2920, 2024, doi: 10.52783/jes.3204.

[14] Đ. Petrović and M. Stanković, "The Influence of Text Preprocessing Methods and Tools on Calculating Text Similarity," *Facta Univ. Ser. Math. Informatics*, p. 973, 2019, doi: 10.22190/fumi1905973d.

[15] C. Caragea, F. Bulgarov, A. Godea, and S. Das Gollapalli, "Citation-Enhanced Keyphrase Extraction From Research Papers: A Supervised Approach," 2014, doi: 10.3115/v1/d14-1150.

[16] B. Trstenjak, S. Mikac, and D. Đonko, "KNN With TF-IDF Based Framework for Text Categorization," *Procedia Eng.*, vol. 69, pp. 1356–1364, 2014, doi: 10.1016/j.proeng.2014.03.129.

[17] D. Hartmann *et al.*, "MISM: A Medical Image Segmentation Metric for Evaluation of Weak Labeled Data," *Diagnostics*, vol. 13, no. 16, p. 2618, 2023, doi: 10.3390/diagnostics13162618.

[18] Y. Tang, L. L. Wen, and G. W. Wei, "Approaches to multiple attribute group decision making based on the generalized Dice similarity measures with intuitionistic fuzzy information," *Int. J. Knowledge-Based Intell. Eng. Syst.*, vol. 21, no. 2, pp. 85–95, 2017, doi: 10.3233/KES-170354.

[19] G. Yunanda, D. Nurjanah, and S. Meliana, "Recommendation System from Microsoft News Data using TF-IDF and Cosine Similarity Methods," *Build. Informatics, Technol. Sci.*, vol. 4, no. 1, pp. 277–284, 2022, doi: 10.47065/bits.v4i1.1670.

[20] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Inf. Sci. (Ny).*, vol. 507, pp. 772–794, 2020, doi: 10.1016/j.ins.2019.06.064.