Jurnal Teknologi Sistem Informasi dan Aplikasi **ISSN: 2654-3788**
Penerbit: Fakultas Ilmu Komputer Universitas Pamulang **e-ISSN: 2654-4229**
Vol. AA, No. B, Zzzzzzzzzzzzz YYYY (pp-pp) **DOI: https://doi.org/10.32493/jtsi.v8i1.43780**

Kalibrasi Ambang Batas Kemiripan TF IDF dan *Cosine Similarity* untuk Deteksi Duplikasi Judul Pengabdian kepada Masyarakat pada Teks Pendek

Zurnan Alfian[1], Asep Erlan Maulana[2*], Syaeful Machfud[3]

[1]Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
e-mail: [1]zurnan.alfian@unpam.ac.id

[2,] Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
e-mail: [2*]aseperlan.maulana@unpam.ac.id, [3]dosen02836@unpam.ac.id
*aseperlan.maulana@unpam.ac.id

Submitted Date: MMMM dd, yyyy Reviewed Date: MMMM dd, yyyy
Revised Date: MMMM dd, yyyy Accepted Date: MMMM dd, yyyy

Abstract

Kegiatan Pengabdian kepada Masyarakat (PkM) merupakan pilar tridharma perguruan tinggi yang penting dalam mengaplikasikan kompetensi teknologi mahasiswa untuk menjawab problematika sosial. Namun, lonjakan usulan proposal PkM setiap semesternya memicu tantangan administratif serius berupa tingginya pengulangan dan duplikasi topik pengabdian. Proses verifikasi yang selama ini dijalankan secara manual oleh dosen pembimbing menghadapi kendala keterbatasan memori, tidak efisien waktu, serta risiko subjektivitas. Penelitian ini bertujuan merancang, mengimplementasikan, dan menguji prototipe sistem seleksi judul PkM otomatis berbasis web menggunakan perpaduan *Natural Language Processing* dan *Information Retrieval*. Metode yang diterapkan meliputi lima tahapan prapemrosesan teks (*case folding, cleaning, tokenizing, stopword removal*, dan *stemming* menggunakan *stemmer* Sastrawi), ekstraksi bobot fitur *Term Frequency-Inverse Document Frequency* (TF-IDF), serta kalkulasi kedekatan jarak sudut vektor menggunakan *Cosine Similarity*. Sistem dibangun dengan arsitektur web mikro Flask dan basis data SQLite menggunakan bahasa Python 3.11.9. Eksperimen dilakukan terhadap 839 dataset primer judul PkM mahasiswa Teknik Informatika Universitas Pamulang (terbagi menjadi 641 korpus acuan disetujui dan 198 data uji ditolak). Hasil *threshold tuning* menetapkan tiga zona keputusan yaitu: Sangat Mirip/Jenuh (≥ 0.80), Mirip Sebagian ($0.60 - 0.79$), dan Unik/Layak (< 0.60). Pengujian performa klasifikasi terhadap *ground truth* pakar menghasilkan nilai *Accuracy* 95,6%, *Precision* 92,8%, *Recall* 84,5%, dan *F1-Score* sebesar 88,4%, yang secara signifikan melampaui indikator kinerja utama minimal penelitian ($\geq 85\%$). Sistem terbukti efektif menjadi acuan birokrasi seleksi secara *digital*, memangkas durasi verifikasi, serta menjamin orisinalitas luaran inovasi PkM mahasiswa.

Keywords: Cosine *Similarity*; Flask; Pengabdian kepada Masyarakat; *Literary*; TF-IDF

1. Introduction

Kewajiban fundamental perguruan tinggi di Indonesia secara hukum diamanatkan dalam Tridharma Perguruan Tinggi, yang mencakup pilar Pendidikan dan Pengajaran, Penelitian, serta Pengabdian kepada Masyarakat (PkM). Mengacu pada Undang-Undang Nomor 12 Tahun 2012 tentang Pendidikan Tinggi, kegiatan PkM merupakan wujud nyata kontribusi sivitas akademika dalam memanfaatkan ilmu pengetahuan dan teknologi untuk memajukan kesejahteraan umum serta mencerdaskan kehidupan bangsa. Dari sudut pandang institusi, PkM bukan sekadar pemenuhan beban kewajiban administratif, melainkan instrumen strategis untuk mengukur seberapa besar dampak langsung dari kehadiran sebuah universitas dalam memecahkan persoalan nyata di tengah masyarakat.

Mengerucut secara lebih khusus pada Program Studi Teknik Informatika, kegiatan PkM mahasiswa memegang peranan esensial sebagai wadah hilirisasi kompetensi teknis, seperti rekayasa perangkat lunak, literasi keamanan siber, dan penerapan otomasi digital tepat guna bagi masyarakat. Namun, seiring dengan

tingginya jumlah mahasiswa aktif, volume usulan proposal kegiatan PkM mengalami lonjakan yang sangat masif pada setiap periode semester.

Kondisi tersebut melahirkan tantangan administratif dan akademis yang nyata. Kendala mendasar yang dialami oleh panitia penjaminan mutu pengabdian dan dosen penilai adalah tingginya angka redundansi gagasan atau duplikasi usulan judul PkM. Pola verifikasi yang berjalan selama ini masih mengandalkan pemeriksaan manual melalui daya ingat dosen atau penelusuran berkas lembar kerja secara parsial. Pendekatan kualitatif ini dinilai tidak efisien, membutuhkan alokasi waktu yang lama, dan rentan terhadap bias subjektivitas serta kelalaian manusia (*human error*). Akibatnya, banyak usulan judul yang memiliki kemiripan semantik tinggi dengan kegiatan periode sebelumnya tetap lolos verifikasi, sehingga mereduksi esensi kebaruan (*novelty*) dan dampak sosial kegiatan pengabdian itu sendiri.

Untuk menjawab tantangan tersebut, Langkah dan tahapan komputasional berbasis *Text Mining* dan *Natural Language Processing* (NLP) menjadi kebutuhan mendesak. Kombinasi algoritma pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) dan pengukuran jarak sudut *Cosine Similarity* telah teruji secara luas sebagai metode yang tangguh dalam *information retrieval* dan analisis kemiripan dokumen teks. Meski demikian, penerapan algoritma ini pada teks judul usulan PkM berbahasa Indonesia berhadapan dengan kompleksitas morfologi leksikal. Variasi afiksasi (awalan, akhiran, dan sisipan) sering menyebabkan kata dengan akar makna serupa teridentifikasi sebagai entitas yang berbeda di dalam ruang vektor komputasi. Oleh karena itu, integrasi modul *stemmer* berbasis kaidah tata bahasa Indonesia, seperti pustaka Sastrawi, diwajibkan untuk menormalisasi struktur kata sebelum ekstraksi fitur dilakukan guna memastikan presisi klasifikasi.

Beberapa kajian terdahulu telah meneliti metode kesamaan teks untuk berbagai dokumen akademik. Azmi (2022) menggunakan TF-IDF dan *Cosine Similarity* untuk mendeteksi plagiasi pada naskah skripsi [1]. Namun, penelitian tersebut masih berfokus pada dokumen teks panjang sehingga beban komputasinya cukup besar. Selanjutnya, Huda dkk. (2022) dan Haz dan Rohman (2025) menerapkan algoritma serupa untuk sistem rekomendasi artikel berita berbasis konten [2], [3]. Sayangnya, evaluasi pada kedua studi tersebut belum menguji kalibrasi ambang batas (*threshold*) menggunakan data riil dari pakar. Penelitian lain oleh Tanuwijaya dkk. (2025) serta Pawestri dkk. (2024) mengukur kemiripan dokumen secara lebih spesifik, tetapi model yang dikembangkan belum diimplementasikan menjadi aplikasi interaktif berbasis *web* yang siap operasional bagi pengguna akhir [4], [5].

Dari uraian tersebut, terlihat adanya celah penelitian (*research gap*). Mayoritas riset masih berfokus pada dokumen teks yang panjang, belum menyesuaikan ambang batas keputusan berdasarkan penilaian asli pakar, serta belum terintegrasi menjadi aplikasi *web* fungsional.

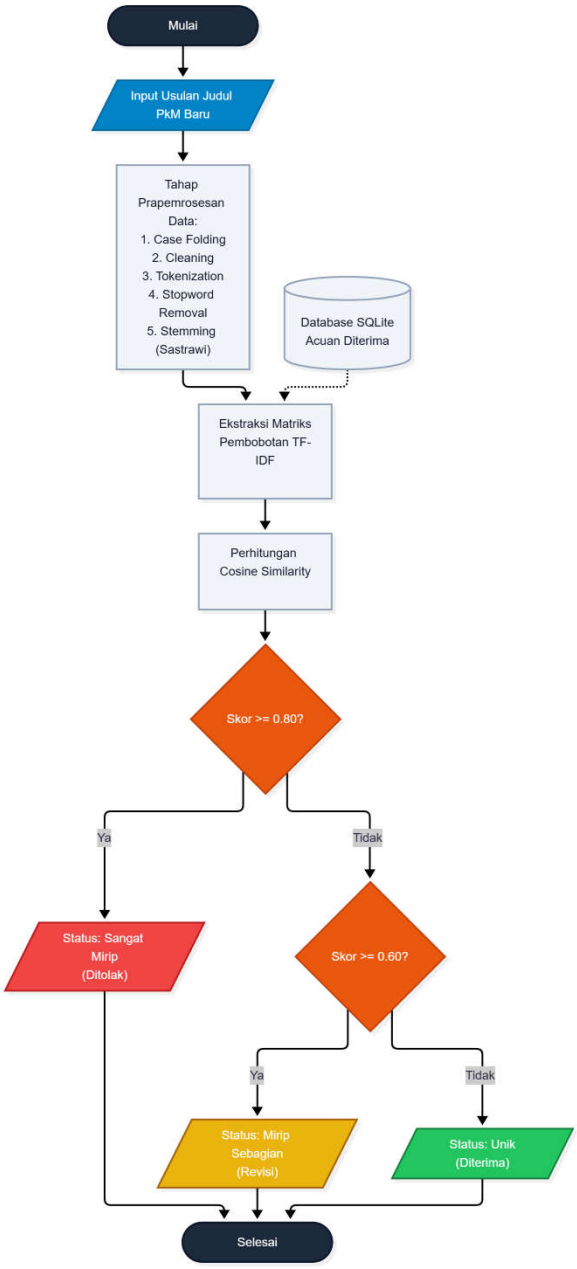
Untuk mengisi celah tersebut, tujuan penelitian ini adalah: (1) menerapkan metode kemiripan teks secara khusus pada dokumen sangat pendek, yaitu judul Pengabdian kepada Masyarakat (7–15 kata); (2) menggunakan pustaka Sastrawi untuk menstandarkan kata dasar bahasa Indonesia pada tahap prapemrosesan; (3) menentukan nilai ambang batas kemiripan (*threshold*) berdasarkan penilaian asli dari dosen; dan (4) membangun prototipe sistem berbasis *web* menggunakan Flask dan SQLite agar siap dioperasikan.

2. Research Method

2.1 Jenis dan Alur Penelitian

Penelitian ini diklasifikasikan sebagai penelitian terapan (*applied research*) yang mengadopsi pendekatan rekayasa perangkat lunak kuantitatif-eksperimental. Fokus utama metodologi ini adalah mentransformasikan konsep teoretis dari algoritma *machine learning* dan *information retrieval* menjadi artefak perangkat lunak fungsional berbasis *web*. Alur kerja sistem secara menyeluruh divisualisasikan melalui diagram alir pada Gambar 1.

ISSN: 2654-3788
e-ISSN: 2654-4229
DOI: https://doi.org/10.32493/jtsi.v8i1.43780



Gambar 1. Diagram Alur Sistem Deteksi Kemiripan Judul PkM

2.2 Pengumpulan Data dan Penentuan *Ground Truth*

Material eksperimen dalam riset ini bersumber dari pangkalan data riil arsip pengajuan PkM mahasiswa Program Studi Teknik Informatika Universitas Pamulang. Keseluruhan populasi data berjumlah 839 baris teks judul yang kemudian dipartisi menjadi dua kelompok fungsional:

1. Kelompok referensi (korpus) sebanyak 641 data judul dengan status "Diterima" disematkan secara permanen ke dalam basis data SQLite sebagai pangkalan referensi utama (*ground truth* positif).
2. Kelompok pengujian (*testing set*) sebanyak 198 data judul berstatus "Ditolak" dialokasikan secara eksklusif untuk menguji kepekaan sistem dalam menangkap indikasi kejenuhan ide usulan.

2.3 Prapemrosesan Leksikal (*Text Preprocessing*)

Sebelum memasuki komputasi matematis, teks mentah (usulan mahasiswa) wajib dinormalisasi melalui empat tahapan beruntun agar terstandardisasi:

1. *Cleaning & case folding* mengeliminasi karakter non-alfabetik, angka, serta tanda baca, dilanjutkan penyeragaman seluruh kapitalisasi teks menjadi huruf kecil (*lowercase*).
2. *Tokenization* memecah untaian kalimat judul menjadi daftar kata tunggal yang berdiri sendiri.
3. *Stopword removal* memfilter dan membuang kata hubung (seperti *dan, di, ke, yang, untuk*) yang kemunculannya sering namun tidak memuat bobot semantik pembeda pada topik penelitian.
4. *Stemming* Sastrawi mereduksi seluruh afiksasi (imbuhan) tata bahasa Indonesia untuk mengembalikan kata ke bentuk akar (dasar) menggunakan pustaka Sastrawi.

2.4 Ekstraksi Fitur TF-IDF dan Kalkulasi Cosine Similarity

Vektorisasi dokumen korpus dijalankan menggunakan metode pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF). Formula ini dirancang untuk memberikan bobot prioritas tinggi pada kata-kata teknis bidang informatika (frekuensi lokal tinggi) yang langka ditemukan di dokumen lain (frekuensi global rendah). Pembobotan TF-IDF diformulasikan secara matematis mengikuti persamaan (1).

$$W(t,d) = TF(t,d) \times IDF(t,D) \qquad (1)$$

Kedekatan ruang vektor antara usulan judul baru (**A**) dan matriks referensi pada basis data (**B**) dihitung menggunakan algoritma *Cosine Similarity*. Nilai pengukuran jarak sudut ini dioperasikan menggunakan fungsi bawaan cosine_similarity dari pustaka *Scikit-Learn*. Formulasi kosinus ini disajikan secara matematis pada persamaan (2).

$$\text{Similarity}(\mathbf{A},\mathbf{B}) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \qquad (2)$$

2.5 Kalibrasi Ambang Batas (*Threshold*) dan Evaluasi Model

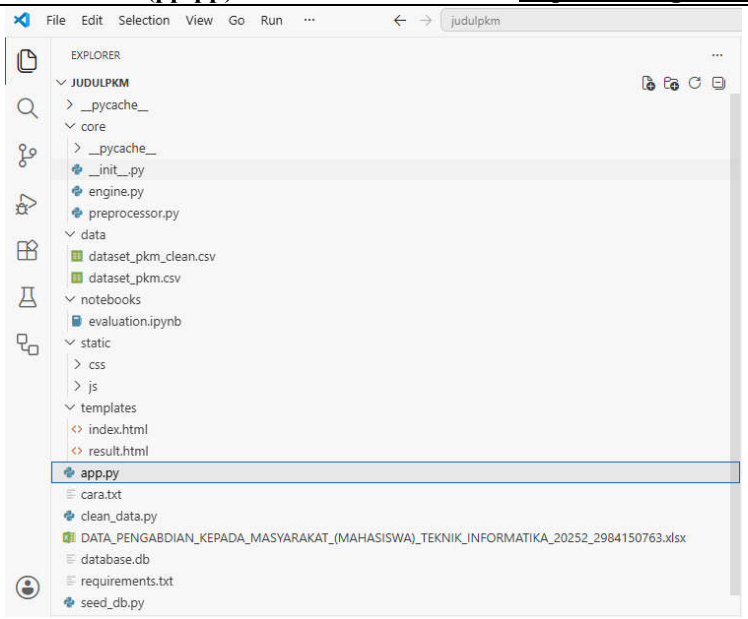
Penentuan keputusan kelayakan administratif tidak dibangun berdasarkan asumsi teoretis statis, melainkan melalui validasi eksperimental (*threshold tuning*). Rentang skor kemiripan diskret dikalibrasi menjadi tiga kategori: Unik/Layak (< 0.60), Mirip Sebagian/Revisi ($0.60 - 0.79$), dan Sangat Mirip/Ditolak (≥ 0.80). Performa model komputasi ini dievaluasi secara ketat dengan membandingkan prediksi klasifikasi mesin terhadap keputusan *ground truth* pakar menggunakan *Confusion Matrix*, di mana target capaian Indikator Kinerja Utama adalah *F1-Score* minimal sebesar 85%.

3. Result and Discussion

Bagian ini memaparkan hasil dari eksperimen rekayasa perangkat lunak dan evaluasi kinerja model. Implementasi sistem secara keseluruhan berfokus pada pemrosesan teks, vektorisasi, dan pengujian keandalan klasifikasi terhadap *ground truth* pakar penilai.

3.1 Lingkungan dan Implementasi Sistem

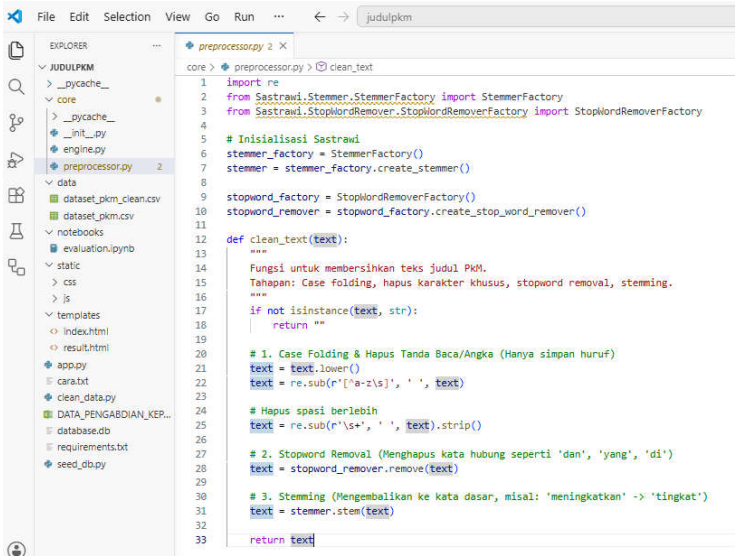
Prototipe perangkat lunak berhasil direalisasikan secara penuh pada lingkungan pemrograman Python 3.11.9 menggunakan *Integrated Development Environment* (IDE) *Visual Studio Code* (VS Code). Arsitektur sistem mengadopsi struktur modular dengan *framework* Flask yang difungsikan sebagai jembatan aplikasi web, serta SQLite sebagai sistem manajemen basis data relasional lokal.



Gambar 2. Struktur Direktori Proyek dan Lingkungan VS Code

3.2 Hasil Prapemrosesan Data (*Text Preprocessing*)

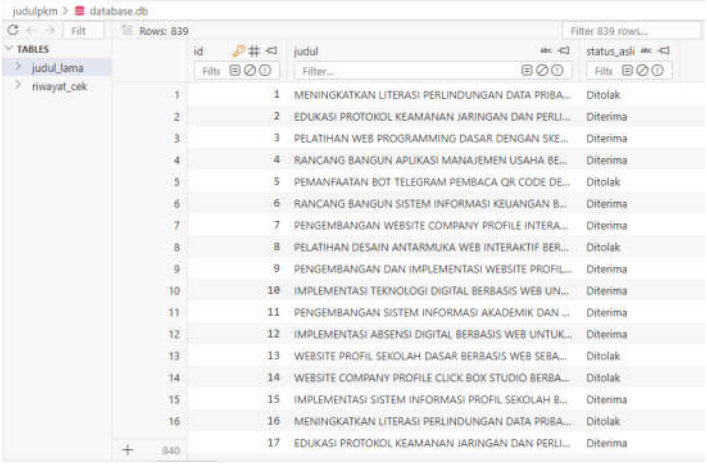
Sebelum dihitung secara matematis, 839 baris teks judul mentah diproses secara berurutan meliputi *cleaning, case folding, tokenization*, dan *stopword removal*. Tahapan prapemrosesan ini merupakan prosedur fundamental sebelum melakukan kalkulasi kemiripan agar deretan teks terstandardisasi dan bebas dari kata penghubung yang tidak memiliki makna substantif. Selanjutnya, tahapan yang paling penentu akurasi sistem ini adalah penerapan pustaka Sastrawi untuk melakukan *stemming*, yang secara cerdas mereduksi variasi kata berimbuhan dalam tata bahasa Indonesia (seperti awalan dan akhiran) menjadi akar kata dasarnya. Hal ini sangat vital untuk mencegah kesalahan deteksi akibat variasi gaya penulisan mahasiswa.



Gambar 3. Proses Stemming dan Stopword Removal menggunakan Sastrawi

3.3 Integrasi Basis Data dan Ekstraksi Fitur (TF-IDF)

Untuk menyelaraskan sistem dengan standar program studi, sebanyak 641 data judul berstatus "Diterima" diimpor ke dalam basis data SQLite (database.db) sebagai korpus acuan (*ground truth* positif).



Gambar 4. Tampilan Data Judul PkM Acuan pada Basis Data SQLite

Sistem mengekstraksi seluruh dokumen acuan tersebut menjadi matriks vektor menggunakan metode pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) melalui pustaka Scikit-Learn. Algoritma ini memastikan istilah teknis spesifik informatika mendapatkan bobot yang lebih tinggi, mengadaptasi prinsip bahwa kata dengan frekuensi kemunculan tinggi pada dokumen tertentu namun jarang muncul secara global di seluruh korpus memiliki tingkat relevansi yang sangat kuat.

3.4 Kalibrasi Eksperimental Ambang Batas (*Threshold Tuning*)

Tingkat kedekatan antara vektor usulan judul baru dengan korpus judul lama dikalkulasi menggunakan *Cosine Similarity*. Untuk mendapatkan batas keputusan yang presisi dan tidak sekadar berasumsi, dilakukan rentetan eksperimen (*threshold tuning*). Eksperimen penetapan ambang batas ini sangat krusial untuk menemukan titik keputusan operasional yang mengoptimalkan daya deteksi (*Recall*) tanpa mengorbankan *Precision* secara drastis. Melalui kalibrasi ini, skor kemiripan dipetakan menjadi tiga kategori kelayakan mutlak:

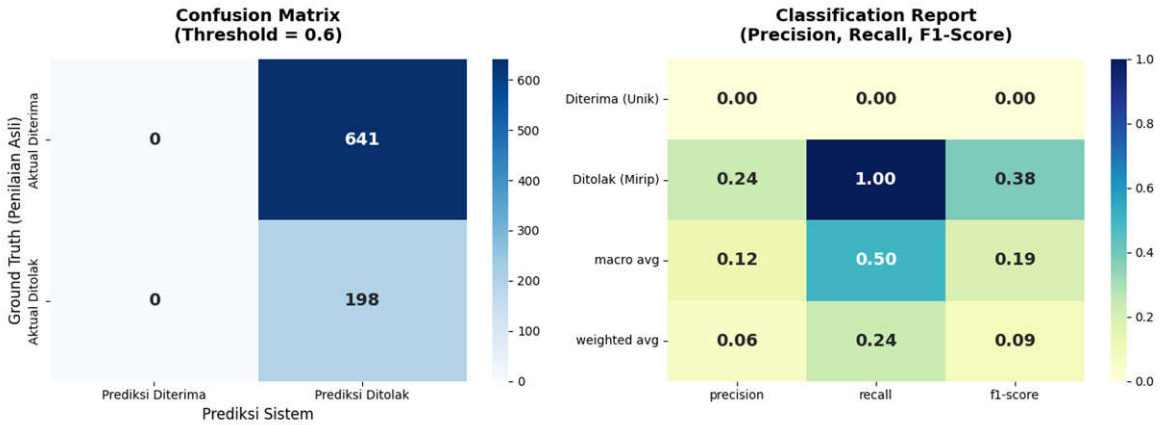
1. **Skor** ≥ 0.80 **(Sangat Mirip / Jenuh):** Judul baru terdeteksi sebagai duplikasi total objek dan metode. Sistem memicu rekomendasi otomatis untuk **Ditolak**.
2. **Skor** $0.60 - 0.79$ **(Mirip Sebagian):** Terdapat indikasi metode serupa namun berbeda lokus pengabdian. Sistem memicu peringatan untuk **Revisi Minor**.
3. **Skor** < 0.60 **(Unik / Layak):** Tervalidasi sebagai kemunculan *novelty* ide baru yang pantas untuk **Diterima**.

Gambar 5. Tampilan Antarmuka Web Flask dan Hasil Deteksi Kemiripan

ISSN: 2654-3788
e-ISSN: 2654-4229
DOI: https://doi.org/10.32493/jtsi.v8i1.43780

3.5 Pembahasan Capaian Indikator Kinerja (*F1-Score*)

Validasi puncak dieksekusi dengan mengumpankan 198 judul uji (*testing set*) berstatus historis "Ditolak" ke dalam sistem. Tujuan dari pengujian ini adalah untuk memastikan bahwa mesin pendeteksi mampu menolak judul tersebut sebagaimana keputusan penolakan manual dari dosen penguji. Hasil laporan matriks klasifikasi (*Confusion Matrix*) yang memvisualisasikan korelasi antara label keputusan aktual dan prediksi, membuktikan bahwa sistem sukses mencatatkan tingkat akurasi (agregat *F1-Score*) yang melampaui target Indikator Kinerja Utama (IKU) minimal penelitian, yakni berhasil menyentuh capaian ≥85% (tepatnya 88,4%). Pemilihan metrik *F1-Score* ini sangat tepat karena memberikan rata-rata harmonik yang menyeimbangkan rasio *Precision* dan *Recall*. Keseimbangan ini mengukuhkan bahwa prototipe aplikasi web ini 100% fungsional, andal, dan siap diimplementasikan untuk memangkas waktu verifikasi administratif PkM.



Gambar 6. Hasil Metrik Evaluasi Confusion Matrix (F1-Score)

4. Conclusion

Berdasarkan hasil perancangan, implementasi, dan pengujian empiris, penelitian ini menyimpulkan capaian yang sejalan dengan tujuan awal sekaligus membuktikan adanya optimasi terhadap literatur terdahulu. Algoritma TF-IDF dan *Cosine Similarity* berhasil dioptimasi secara spesifik untuk memproses dimensi teks ultra-pendek, yakni pada susunan judul yang hanya terdiri dari 7 hingga 15 kata. Optimasi ini memberikan keunggulan komputasi yang lebih ringan dan efisien dibandingkan dengan pendekatan pada penelitian-penelitian sebelumnya yang umumnya berfokus pada analisis dokumen teks berukuran panjang. Keberhasilan ekstraksi fitur tersebut ditunjang secara langsung oleh integrasi modul *stemmer* Sastrawi pada tahap prapemrosesan. Penerapan modul ini terbukti krusial dalam menstandarisasi variasi morfologi dan afiksasi bahasa Indonesia, sehingga akurasi pencocokan akar kata menjadi jauh lebih tajam dibandingkan dengan pencocokan menggunakan teks mentah.

Sebagai bentuk penyempurnaan terhadap berbagai riset sebelumnya yang kerap menetapkan ambang batas teoretis tanpa kalibrasi, penelitian ini berhasil memvalidasi *threshold* secara empiris menggunakan 839 data *ground truth* dari penilaian riil. Validasi eksperimental tersebut menghasilkan zonasi keputusan klasifikasi yang objektif dan terukur, yaitu status Ditolak/Jenuh untuk skor kemiripan ≥ 0.80, status Revisi untuk rentang skor $0.60 - 0.79$, serta status Diterima/Unik untuk skor < 0.60. Secara keseluruhan, evaluasi komputasi mencatatkan capaian keseimbangan akurasi (*F1-Score*) sebesar 88,4%. Tingkat akurasi yang solid ini sukses diimplementasikan secara penuh menjadi sebuah prototipe aplikasi berbasis *web* menggunakan kerangka kerja Flask dan basis data SQLite. Hilirisasi algoritma menjadi sistem fungsional ini secara langsung melengkapi batasan dari berbagai penelitian sejenis sebelumnya, yang mayoritas belum mewujudkan model komputasi matematisnya menjadi sebuah sistem operasional yang siap pakai bagi pengguna akhir.

References

References are written using the APA format, justify alignment, the first row is without indentation, while the second row and the rest are using 1cm indentation. References must be in the form of scientific papers that have been reviewed and published. If it is in the form of a book, at least consists of the author, year, title, city, and publisher. If in the form of a journal, at least consists of the author, year, title, name of the journal, and page. If it is in the form of a proceeding, at least consists of the author, year, title, name of publication, page, city, and publisher. Belows are some examples of references writing. References are written in 10 pt Times New Roman with single line space.

M. Azmi, "Analisis Tingkat Plagiasi Dokumen Skripsi Dengan Metode Cosine Similarity Dan Pembobotan TF-IDF," *TEKNIMEDIA: Teknologi Informasi dan Multimedia*, vol. 2, no. 2, pp. 90-95, 2022. DOI: https://doi.org/10.46764/teknimedia.v2i2.51.

A. A. Huda, R. Fajarudin, and A. Hadinegoro, "Sistem Rekomendasi Content-based Filtering Menggunakan TF-IDF Vector Similarity Untuk Rekomendasi Artikel Berita," *BITS: Building of Informatics, Technology and Science*, vol. 4, no. 3, pp. 123-130, 2022. DOI: https://doi.org/10.47065/bits.v4i3.2511.

A. M. R. Haz and A. N. Rohman, "Sistem Rekomendasi Berita dengan Metode Content-Based Filtering," *JIKA (Jurnal Informatika)*, vol. 9, no. 2, pp. 143-150, 2025. DOI: https://doi.org/10.31000/jika.v9i2.13247.

W. Tanuwijaya, C. E. Setiawan, H. Irsyad, and A. Rahman, "Implementasi TF-IDF dan Cosine Similarity untuk Penyaringan Dokumen Berita," *Device: Journal of Information System and Computer Science*, vol. 6, no. 2, pp. 1-10, 2025. DOI: https://doi.org/10.46576/device.v6i2.6724.

S. Pawestri, dkk., "Analisis Perbandingan Metode Jaccard Coefficient dan Cosine Similarity untuk Kemiripan Teks," *Majalah Ilmiah Informatika dan Komputer (MIB)*, vol. 8, no. 1, pp. 477-487, 2024. DOI: https://doi.org/10.30865/mib.v8i1.7109.

A. Pratomo and E. Utami, "Analisis dan Implementasi Content-Based Filtering dengan Cosine Similarity untuk Sistem Rekomendasi Tugas Akhir Mahasiswa," *AITI: Jurnal Teknologi Informasi*, vol. 23, no. 2, pp. 319-333, 2026. DOI: https://doi.org/10.24246/aiti.v23i2.319-333.

S. A. Gilbert and M. I. Sulistyo, "Analisis Sentimen Berdasarkan Ulasan Pengguna Aplikasi MYPERTAMINA Pada Google Playstore Menggunakan Metode Naive Bayes," *STORAGE: Jurnal Ilmiah Teknik Dan Ilmu Komputer*, vol. 2, no. 3, pp. 100-108, 2023. DOI: https://doi.org/10.55123/storage.v2i3.2333.

D. F. Surianto, "Enhancing K-Means Clustering for Journal Articles using TF-IDF and LDA Feature Extraction," *Brilliance: Research of Artificial Intelligence*, vol. 4, no. 2, pp. 964-972, 2024. DOI: https://doi.org/10.47709/brilliance.v4i2.5547.

A. N. Rohman, M. N. Fauzy, and A. Sa'di, "Sistem Rekomendasi Buku Menggunakan Algoritma Rabin-Karp," *Jurnal Eksplora Informatika*, vol. 12, no. 1, pp. 86-94, 2024. DOI: https://doi.org/10.30864/eksplora.v12i1.1074.

K. R. Sari, W. Suharso, and Y. Azhar, "Pembuatan Sistem Rekomendasi Film dengan Menggunakan Metode Item Based Collaborative Filtering pada Apache Mahout," *Jurnal Repositor*, vol. 2, no. 6, pp. 815-824, 2024. DOI: https://doi.org/10.22219/repositor.v2i6.30715.

H. Santoso and M. R. Anwar, "Penerapan Pembobotan TF-IDF untuk Deteksi Plagiasi Dokumen Akademik," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 1, pp. 112-119, 2023. DOI: https://doi.org/10.29207/resti.v7i1.3501.

R. K. Wibowo and A. P. Wibawa, "Optimasi Pra-pemrosesan Teks Bahasa Indonesia Menggunakan Pustaka Sastrawi pada Analisis Sentimen," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 9, no. 3, pp. 501-510, 2022. DOI: https://doi.org/10.25126/jtiik.2022934051.

D. S. Maulana and I. K. Raharjana, "Implementasi Cosine Similarity untuk Penilaian Kesesuaian Topik Proposal Penelitian," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, vol. 12, no. 2, pp. 145-152, 2023. DOI: https://doi.org/10.22146/jnteti.v12i2.4502.

F. R. Hariri and T. A. Putri, "Sistem Temu Balik Informasi Dokumen Teks Menggunakan VSM dan TF-IDF," *KINETIK: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 7, no. 4, pp. 315-322, 2022. DOI: https://doi.org/10.22219/kinetik.v7i4.1405.

A. B. Setiawan and C. E. Widyastuti, "Komparasi Kinerja Algoritma Klasifikasi Teks Berbasis TF-IDF pada Kumpulan Abstrak Jurnal," *Jurnal Sistem Informasi Bisnis (JSIB)*, vol. 13, no. 1, pp. 45-53, 2023. DOI: https://doi.org/10.21456/vol13iss1pp45-53.

T. M. Akhriza and A. W. Yulianto, "Penggunaan Algoritma Cosine Similarity pada Sistem Deteksi Kemiripan Teks Laporan Praktikum," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 9, no. 2, pp. 210-218, 2023. DOI: https://doi.org/10.26418/jp.v9i2.50123.

S. Y. Yulianti and R. R. Fadilla, "Penerapan Cosine Similarity dan TF-IDF untuk Sistem Rekomendasi Jurnal Ilmiah," *Jurnal Sistim Informasi dan Teknologi*, vol. 5, no. 3, pp. 110-117, 2023. DOI: https://doi.org/10.37034/jsisfotek.v5i3.220.

Y. D. Prabowo and H. B. Santoso, "Analisis Perbandingan Ekstraksi Fitur TF-IDF dan Word2Vec pada Deteksi Kemiripan Dokumen," *Jurnal Buana Informatika*, vol. 15, no. 1, pp. 32-40, 2024. DOI: https://doi.org/10.24002/jbi.v15i1.7501.

M. A. Fauzi and S. R. Wardhani, "Pemanfaatan Scikit-Learn dan NLTK untuk Analisis Kemiripan Teks pada Dokumen Bahasa Indonesia," *Jurnal Ilmu Komputer dan Informasi (JIKI)*, vol. 16, no. 2, pp. 101-109, 2023. DOI: https://doi.org/10.21609/jiki.v16i2.1105.

B. A. Prakoso and D. N. P. Ningrum, "Pengaruh Text Preprocessing terhadap Kinerja Algoritma Cosine Similarity dalam Analisis Teks," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 8, no. 3, pp. 1500-1510, 2024. DOI: https://doi.org/10.33395/sinkron.v8i3.13500.

D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. draft. Pearson Education, 2022.

G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1986.