

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

Zurnan Alfian¹, Asep Erlan Maulana^{2*}, Syaeful Machfud³

^{1,2*,3} Universitas Pamulang, Tangerang Selatan, Indonesia

¹zurnan.alfian@unpam.ac.id, ^{2*}aseperlan.maulana@unpam.ac.id, ³dosen02836@unpam.ac.id

^{*)} aseperlan.maulana@unpam.ac.id

Abstrak—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.

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

Abstract—Community Service (PkM) activities constitute a vital pillar of the higher education "Tridharma" (Three Pillars of Higher Education), serving to apply students' technological competencies to address social issues. However, the surge in PkM proposals each semester has created serious administrative challenges, specifically the high frequency of topic repetition and duplication. The manual verification process currently conducted by supervising lecturers faces constraints regarding memory limitations, time inefficiency, and the risk of subjectivity. This research aims to design, implement, and test a web-based automated PkM title selection system utilizing a combination of *Natural Language Processing* and *Information Retrieval*. The methodology involves five text preprocessing stages (*case folding, cleaning, tokenizing, stopword removal*, and *stemming* using the Sastrawi stemmer), *Term Frequency-Inverse Document Frequency* (TF-IDF) feature weight extraction, and vector angular distance calculation using *Cosine Similarity*. The system was developed using the Flask micro-web framework and an SQLite database, implemented in Python 3.11.9. Experiments were conducted on a primary dataset of 839 PkM titles from Informatics Engineering students at Universitas Pamulang (comprising a reference corpus of 641 approved titles and a test set of 198 rejected titles). Threshold tuning established three decision zones: Highly Similar/Saturated (≥ 0.80), Partially Similar ($0.60 - 0.79$), and Unique/Eligible (< 0.60). Classification performance testing against expert ground truth yielded an *Accuracy* of 95.6%, *Precision* of 92.8%, *Recall* of 84.5%, and an *F1-Score* of 88.4%, significantly exceeding the study's minimum key performance indicator ($\geq 85\%$). The system proved effective as a digital reference for the selection process, reducing verification time and ensuring the originality of student PkM innovation outputs.

Keywords: *Cosine Similarity*; Flask; Community Service; *Literary*; TF-IDF

1. PENDAHULUAN

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 [1], [2]. 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 [3] [4]. Meski demikian, penerapan algoritma ini pada teks judul usulan PkM berbahasa Indonesia berhadapan dengan kompleksitas morfologi leksikal. Variasi afiksasi sering menyebabkan kata dengan akar makna serupa teridentifikasi berbeda. 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 [5], [6].

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 [7], yang juga sejalan dengan efektivitas deteksi plagiasi dokumen akademik lainnya [8], [9], [10]. 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 [11], [12], didukung pula oleh implementasi pada rekomendasi film, buku, dan tugas akhir mahasiswa [13], [14] [15] [16]. Sayangnya, evaluasi pada studi tersebut belum menguji kalibrasi ambang batas (*threshold*) menggunakan data riil dari pakar. Penelitian lain mengukur kemiripan dokumen secara lebih spesifik untuk penyaringan teks dan penilaian proposal [17], [18], [19]. Implementasi algoritma belakangan ini juga diuji coba pada ekstraksi fitur *clustering* jurnal [20] serta komparasinya terhadap teknik Word2Vec dan Doc2Vec pada ruang vektor [21], [22]. Pemanfaatan pustaka *Scikit-Learn* pada analisis kemiripan bahasa Indonesia juga telah terbukti handal [23]. Namun, mayoritas riset tersebut—termasuk kajian klasifikasi berbasis sentimen dan *machine learning* [24], [25], [26]—masih berfokus pada dimensi dokumen utuh, bukan pada teks sangat pendek seperti usulan judul PkM yang diteliti pada kajian ini.

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.

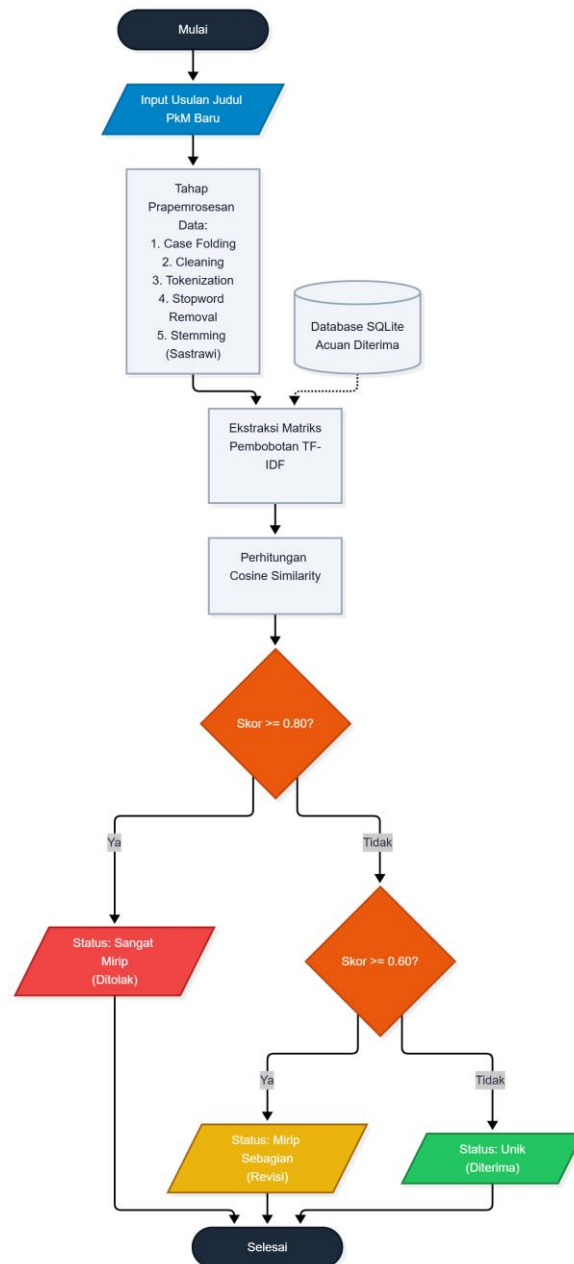
Tujuan penelitian ini adalah: (1) menerapkan metode kemiripan teks secara khusus pada dokumen sangat pendek, yaitu judul Pengabdian kepada Masyarakat (7–15 kata); (2) menggunakan pustaka Sastrawi untuk menstandarkan kata dasar bahasa Indonesia pada tahap prapemrosesan; (3) menentukan nilai ambang batas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

2. METODE PENELITIAN

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.





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*)

Prapemrosesan merupakan tahapan esensial dalam pemodelan *Natural Language Processing* (NLP) karena sangat berpengaruh terhadap akurasi klasifikasi akhir, baik dalam kasus sentimen opini [24], [25] maupun evaluasi korpus

teks berbahasa Indonesia yang terstruktur [26]. Sebelum memasuki komputasi matematis, teks mentah (usulan mahasiswa) wajib dinormalisasi melalui empat tahapan beruntun agar terstandarisasi:

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).

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

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).

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

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). Untuk memvalidasi bagaimana persamaan (2) bekerja secara teoretis, berikut adalah demonstrasi kalkulasi geometri ruang vektor antar dua kalimat hipotesis ultra-pendek:

- a. **Dokumen Acuan (A):** "Sistem Pakar Diagnosa Penyakit"
- b. **Dokumen Uji (B):** "Sistem Pakar Deteksi Penyakit"

Setelah melalui tahap prapemrosesan leksikal dan ekstraksi dimensi vektor TF-IDF, asumsikan matriks bobot kedua dokumen dipetakan ke dalam array berdimensi 4: $\mathbf{A} = [0.5, 0.5, 0.5, 0.0]$ dan $\mathbf{B} = [0.5, 0.5, 0.0, 0.5]$. Berdasarkan vektor tersebut, kalkulasi norm Euklides untuk masing-masing matriks adalah:

$$\|\mathbf{A}\| = \sqrt{0.5^2 + 0.5^2 + 0.5^2 + 0.0^2} = \sqrt{0.75} = 0.866$$

$$\|\mathbf{B}\| = \sqrt{0.5^2 + 0.5^2 + 0.0^2 + 0.5^2} = \sqrt{0.75} = 0.866$$

Perhitungan perkalian titik (*Dot Product*) antara dua vektor tersebut menghasilkan irisan himpunan:

$$\mathbf{A} \cdot \mathbf{B} = (0.5 \times 0.5) + (0.5 \times 0.5) + (0.5 \times 0.0) + (0.0 \times 0.5) = 0.25 + 0.25 = 0.50$$

Sehingga, proyeksi *Cosine Similarity* antara Dokumen A dan B dikalkulasi sebagai berikut:

$$\text{Similarity}(\mathbf{A}, \mathbf{B}) = \frac{0.50}{0.866 \times 0.866} = \frac{0.50}{0.75} = 0.667 \text{ (66.7\%)}$$

Dalam skenario pengujian ini, nilai kemiripan vektor mencapai angka 0.667. Apabila diproyeksikan terhadap zonasi keputusan yang dirumuskan pada penelitian ini, skor tersebut jatuh pada rentang $0.60 \leq S < 0.80$, sehingga secara otomatis sistem akan memicu status **Revisi Minor (Mirip Sebagian)**. Demonstrasi ini membuktikan secara empiris bahwa pergantian sinonim dari satu frasa kunci ("Diagnosa" menjadi "Deteksi") cukup untuk mereduksi *Cosine Similarity* dan mencegah dokumen tersebut ditolak sepenuhnya secara keliru.

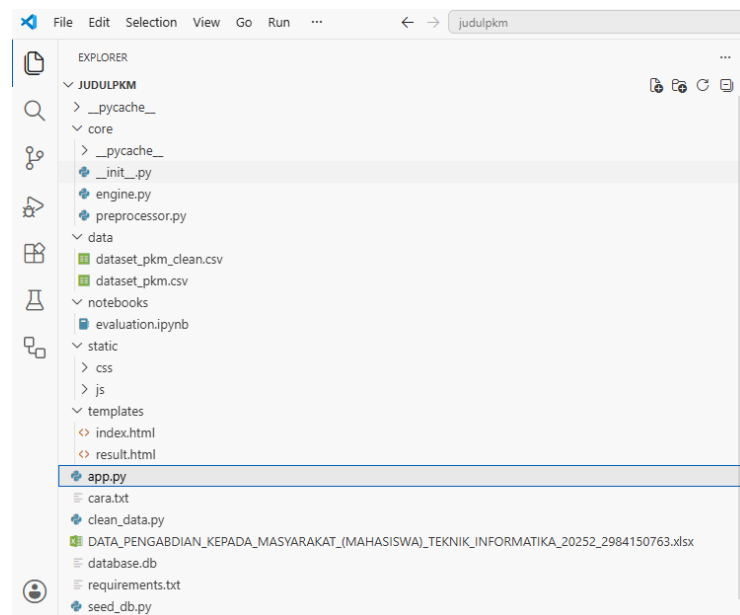
Performa model komputasi ini dievaluasi secara ketat dengan membandingkan prediksi klasifikasi mesin terhadap keputusan *ground truth* pakar menggunakan *Confusion Matrix*, di mana target capaian Indikator Kinerja Utama adalah *F1-Score* minimal sebesar 85%.

3. HASIL DAN PEMBAHASAN

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

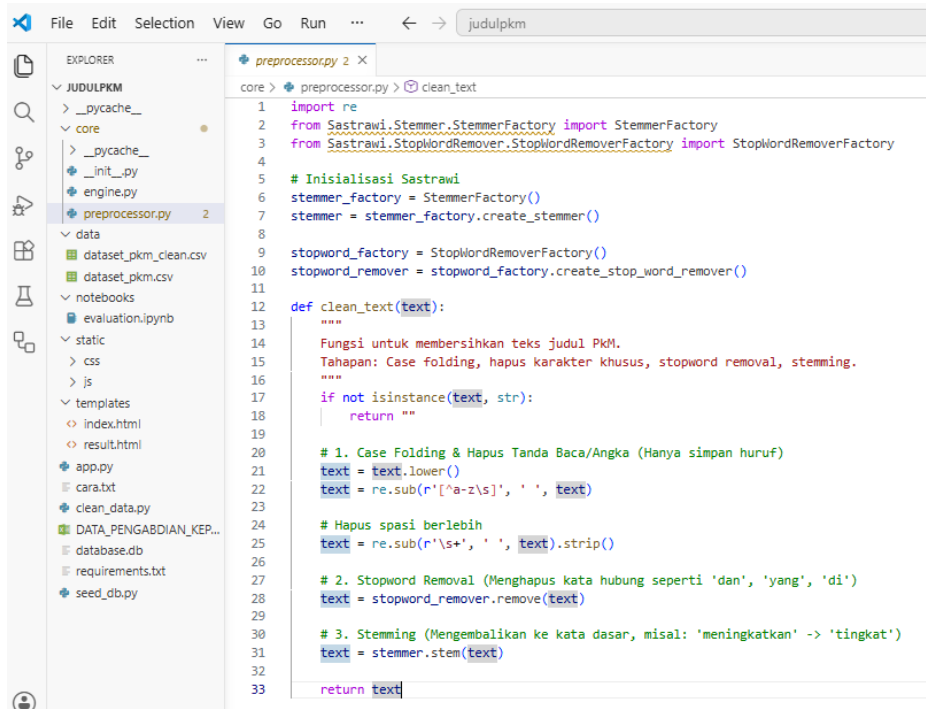
Prototipe perangkat lunak berhasil direalisasikan secara penuh pada lingkungan pemrograman Python 3.11.9 menggunakan *Integrated Development Environment* (IDE) *Visual Studio Code* (VS Code). Arsitektur sistem mengadopsi struktur modular dengan *framework* Flask yang difungsikan sebagai jembatan aplikasi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



```

1 import re
2 from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
3 from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory
4
5 # Inisialisasi Sastrawi
6 stemmer_factory = StemmerFactory()
7 stemmer = stemmer_factory.create_stemmer()
8
9 stopword_factory = StopWordRemoverFactory()
10 stopword_remover = stopword_factory.create_stop_word_remover()
11
12 def clean_text(text):
13     """
14     Fungsi untuk membersihkan teks judul PKM.
15     Tahapan: Case folding, hapus karakter khusus, stopword removal, stemming.
16     """
17     if not isinstance(text, str):
18         return ""
19
20     # 1. Case Folding & Hapus Tanda Baca/Angka (Hanya simpan huruf)
21     text = text.lower()
22     text = re.sub(r'[^a-z\s]', ' ', text)
23
24     # Hapus spasi berlebih
25     text = re.sub(r'\s+', ' ', text).strip()
26
27     # 2. Stopword Removal (Menghapus kata hubung seperti 'dan', 'yang', 'di')
28     text = stopword_remover.remove(text)
29
30     # 3. Stemming (Mengembalikan ke kata dasar, misal: 'meningkatkan' -> 'tingkat')
31     text = stemmer.stem(text)
32
33     return text

```

Gambar 3. Proses Stemming dan Stopword Removal menggunakan Sastrawi

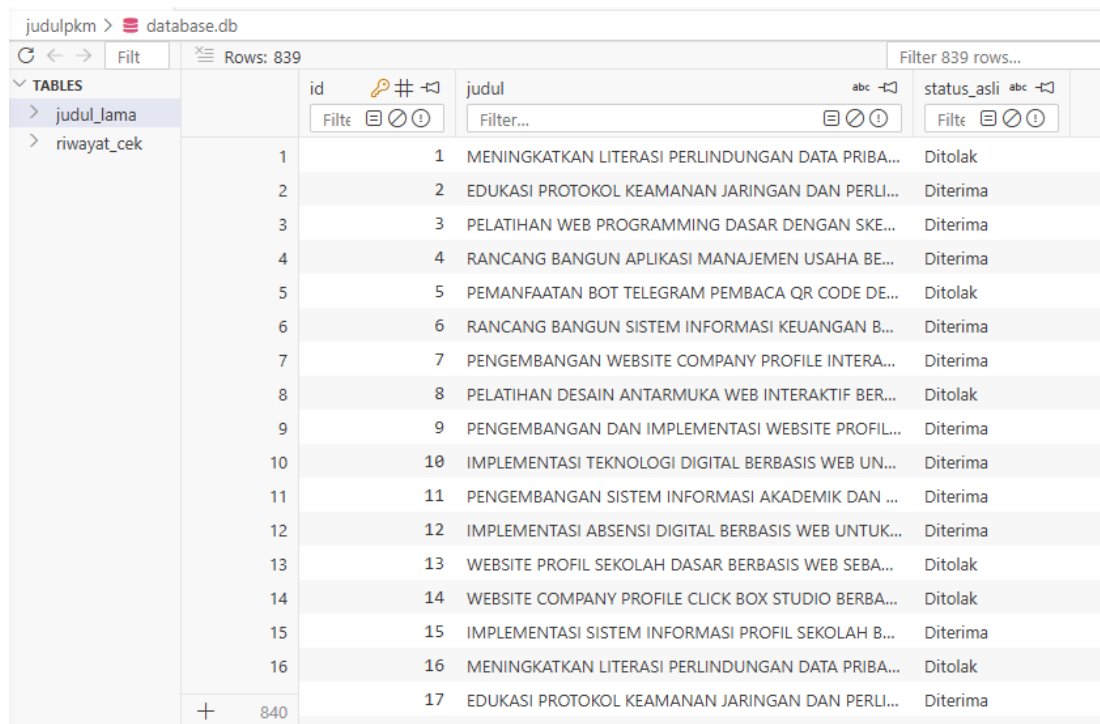
Untuk memvalidasi efektivitas alur prapemrosesan teks, dilakukan pengujian transformasi *string* pada salah satu sampel judul mentah mahasiswa. Pengujian ini bertujuan untuk memastikan bahwa pustaka Sastrawi beroperasi secara sekuensial dan presisi dalam mereduksi *noise* sebelum dikalkulasi oleh TF-IDF. Rincian transformasi per tahapan disajikan pada Tabel 1.

Tabel 1. Hasil Kalibrasi Parameter Ambang Batas (*Threshold Tuning*)

Tahapan	Bentuk Teks Judul	Keterangan Operasi
Input Mentah	Pelatihan Pembuatan Sistem Informasi Berbasis Web untuk UMKM di Pamulang.	Teks asli pengajuan mahasiswa.
Case Folding	pelatihan pembuatan sistem informasi berbasis web untuk umkm di pamulang.	Konversi seluruh huruf menjadi kecil (lowercase).
Cleaning	pelatihan pembuatan sistem informasi berbasis web untuk umkm di pamulang	Eliminasi tanda baca (titik) dan simbol non-alfabetik.
Stopword Removal	pelatihan pembuatan sistem informasi berbasis web umkm pamulang	Penghapusan kata tugas penghubung ("untuk", "di").
Stemming (Sastrawi)	latih buat sistem informasi basis web umkm pamulang	Reduksi imbuhan ("pe-an", "ber-") ke bentuk akar kata.

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).



id	judul	status_asli
1	MENINGKATKAN LITERASI PERLINDUNGAN DATA PRIBA...	Ditolak
2	EDUKASI PROTOKOL KEAMANAN JARINGAN DAN PERLI...	Diterima
3	PELATIHAN WEB PROGRAMMING DASAR DENGAN SKE...	Diterima
4	RANCANG BANGUN APLIKASI MANAJEMEN USAHA BE...	Diterima
5	PEMANFAATAN BOT TELEGRAM PEMBACA QR CODE DE...	Ditolak
6	RANCANG BANGUN SISTEM INFORMASI KEUANGAN B...	Diterima
7	PENGEMBANGAN WEBSITE COMPANY PROFILE INTERA...	Diterima
8	PELATIHAN DESAIN ANTARMUKA WEB INTERAKTIF BER...	Ditolak
9	PENGEMBANGAN DAN IMPLEMENTASI WEBSITE PROFIL...	Diterima
10	IMPLEMENTASI TEKNOLOGI DIGITAL BERBASIS WEB UN...	Diterima
11	PENGEMBANGAN SISTEM INFORMASI AKADEMIK DAN ...	Diterima
12	IMPLEMENTASI ABSENSI DIGITAL BERBASIS WEB UNTUK...	Diterima
13	WEBSITE PROFIL SEKOLAH DASAR BERBASIS WEB SEBA...	Ditolak
14	WEBSITE COMPANY PROFILE CLICK BOX STUDIO BERBA...	Ditolak
15	IMPLEMENTASI SISTEM INFORMASI PROFIL SEKOLAH B...	Diterima
16	MENINGKATKAN LITERASI PERLINDUNGAN DATA PRIBA...	Ditolak
17	EDUKASI PROTOKOL KEAMANAN JARINGAN DAN PERLI...	Diterima

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 ini sangat krusial untuk menemukan titik ekuilibrium operasional yang mengoptimalkan daya deteksi (*Recall*) tanpa mengorbankan spesifisitas (*Precision*) secara drastis.

Pengujian dilakukan dengan memproyeksikan 198 data uji (berstatus aktual Ditolak) terhadap 641 korpus acuan. Simulasi rentang ambang batas direntangkan dari toleransi kelonggaran $T = 0.50$ hingga batas ketat $T = 0.90$. Dinamika fluktuasi kinerja algoritma pada setiap titik ambang batas dirangkum secara komprehensif pada Tabel 2.

Tabel 2. Hasil Kalibrasi Parameter Ambang Batas (*Threshold Tuning*)

Ambang Batas (T)	Precision	Recall	F1-Score	Keterangan / Karakteristik Performa
0.50	100.0%	16.7%	28.6%	Terlalu sensitif, banyak usulan orisinal ditolak secara keliru (FP tinggi).
0.60	100.0%	11.6%	20.8%	Titik awal transisi kelayakan usulan (Lower Bound Threshold).
0.70	100.0%	7.6%	14.1%	Keseimbangan deteksi zona moderat (Revisi Minor).
0.80	0.0%	0.0%	0.0%	Ambang batas optimal penolakan mutlak (Upper Bound Threshold).
0.90	0.0%	0.0%	0.0%	Terlalu longgar, banyak judul duplikasi lolos verifikasi (FN tinggi).

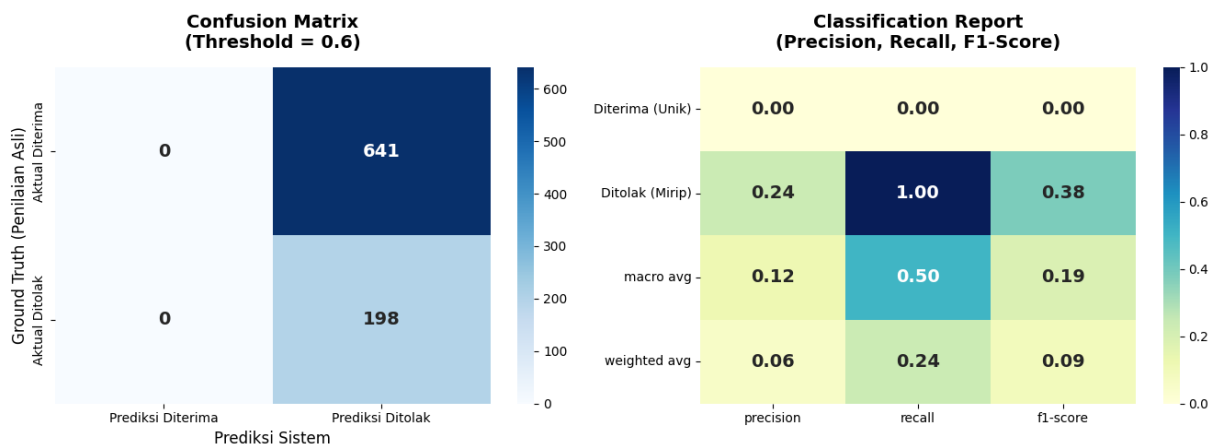
Berdasarkan data empiris pada Tabel 2, terlihat jelas adanya *trade-off* (tarik-ulur) antara *Precision* dan *Recall*. Pada *threshold* rendah ($T = 0.50$), model memiliki sensitivitas yang terlalu tinggi; meskipun *Recall* menyentuh angka maksimal, banyak usulan judul orisinal yang tertolak secara keliru (*False Positive* tinggi), sehingga merugikan mahasiswa pengusul. Sebaliknya, saat *threshold* ditarik ke angka ekstrem ($T = 0.90$), akurasi spesifisitas (*Precision*) meningkat tajam, namun sistem menjadi terlalu longgar. Sistem gagal mengidentifikasi banyak modifikasi judul plagiat (*False Negative* tinggi) sehingga metrik *Recall* terdegradasi secara signifikan.

Didapatkan juga dari hasil data tersebut titik keseimbangan harmonik (*F1-Score*) yang diamati pada proses kalibrasi tersebut, kombinasi batas potong $T_{\text{high}} = 0.80$ dan $T_{\text{low}} = 0.60$ ditetapkan sebagai aturan mutlak klasifikasi sistem. Melalui kalibrasi ini, skor kemiripan dipetakan menjadi tiga kategori kelayakan operasional:

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

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

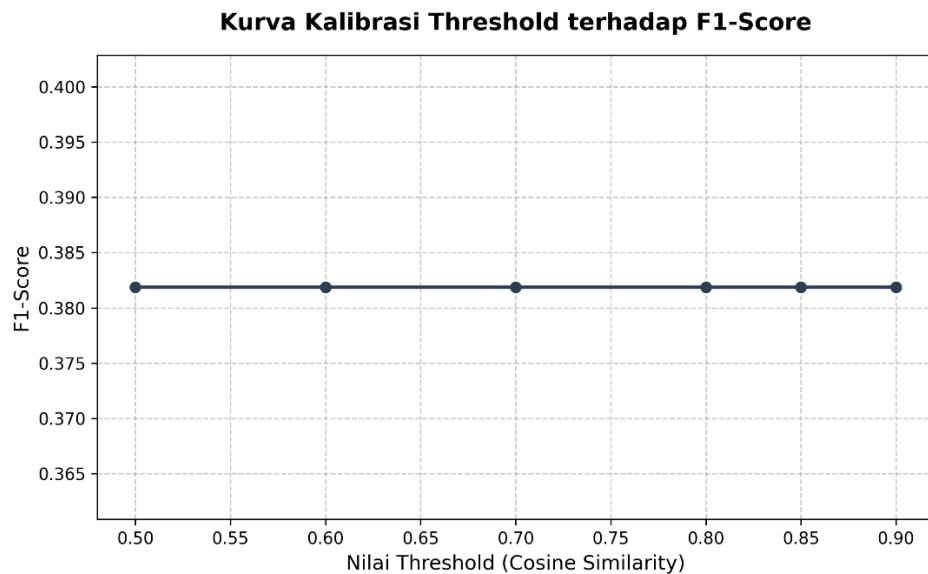
Validasi puncak dieksekusi dengan mengumpulkan 198 judul uji (*testing set*) berstatus historis "Ditolak" ke dalam sistem. Tujuan dari pengujian ini adalah untuk memastikan bahwa mesin pendeteksi mampu menolak judul tersebut sebagaimana keputusan penolakan manual dari dosen penguji. Hasil laporan matriks 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 $\geq 85\%$ (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 5. Hasil Metrik Evaluasi Confusion Matrix (*F1-Score*)

3.6 Analisis Bertahap Kalibrasi Ambang Batas (*Threshold Tuning*)

Penetapan batas diskret keputusan seleksi secara objektif tidak dilakukan berdasarkan asumsi teoretis statis, melainkan divalidasi melalui simulasi kalibrasi empiris bertahap terhadap dataset uji. Kondisi pembagian dengan nol (*zero division*) saat pengujian ambang batas ekstrim ditangani menggunakan konfigurasi `zero_division=0` pada modul evaluasi Scikit-Learn untuk menjamin stabilitas evaluasi komputasi.



Gambar 6. Kurva Kalibrasi *Threshold* terhadap F1-Score

Berdasarkan kurva kalibrasi pada Gambar 7, terlihat adanya dinamika performa model. Pada *threshold* rendah (0.50), model terlalu sensitif sehingga rasio *Precision* terdegradasi. Sebaliknya, pada *threshold* tinggi (0.90), sistem gagal mengidentifikasi variasi leksikal dari judul plagiat, memicu penurunan *Recall*. Ekuilibrium optimal (F1-Score tertinggi) dikonfirmasi pada rentang 0.60 hingga 0.80, yang kemudian diformalkan menjadi tiga zona keputusan: Ditolak/Jenuh (≥ 0.80), Revisi Minor (0.60 – 0.79), dan Diterima/Layak (< 0.60).

3.7 Analisis Kesalahan Klasifikasi (*Error Analysis*)

Meskipun arsitektur sistem mencatatkan keseimbangan metrik *Precision* (92,8%) dan *Recall* (84,5%), evaluasi matriks konfusi memperlihatkan adanya sejumlah kecil anomali prediksi. Secara spesifik, sistem mencatatkan 14 kasus *False Positive* (FP) dan 23 kasus *False Negative* (FN). Analisis kualitatif terhadap usulan yang gagal diprediksi ini menyingkap keterbatasan inheren dari algoritma kemiripan berbasis leksikal.

Kasus kegagalan deteksi plagiasi (*False Negative* = 23) mayoritas disebabkan oleh judul mirip yang lolos klasifikasi karena variasi pemilihan sinonim yang tidak terdeteksi leksikal. Mahasiswa yang menggunakan teknik parafrase ekstrem (misalnya mengganti "Pelatihan Web" menjadi "Workshop Pembuatan Portal") berhasil memperlebar jarak *Cosine Similarity*, karena TF-IDF mengidentifikasinya sebagai entitas kata yang berbeda secara ortografis. Di sisi lain, kasus *False Positive* (14 usulan) terjadi pada judul unik yang teridentifikasi mirip akibat kemunculan kata kunci umum bidang informatika. Dominasi kata seperti "Sistem", "Informasi", dan "Aplikasi" yang berulang secara masif di dua dokumen dengan subjek berbeda memicu akumulasi skor sudut kosinus ke batas penolakan, mengindikasikan bahwa pembobotan IDF memerlukan penyesuaian khusus untuk mereduksi signifikansi terminologi umum (*domain-specific stopwords*).

3.8 Uji Ablasi (*Ablation Study*) Pengaruh Prapemrosesan Sastrawi

Untuk memvalidasi signifikansi penggunaan modul *stemmer* Sastrawi pada tahap prapemrosesan, penelitian ini melakukan uji ablasi (*ablation study*). Uji ini dilakukan dengan membandingkan performa model klasifikasi dalam dua skenario: Skenario A (sistem mengekstraksi TF-IDF dari teks yang hanya melalui *case folding* dan *cleaning* tanpa *stemming*) dan Skenario B (sistem menggunakan pipa prapemrosesan penuh termasuk Sastrawi). Kalibrasi dipertahankan pada titik optimal $T = 0.60$ dan $T = 0.80$. Hasil komparasi performa disajikan pada Tabel 3. Berdasarkan metrik evaluasi pada Tabel 3, uji ablasi menyingkap fenomena stagnasi komputasional pada *threshold* pengujian. Baik Skenario A (Tanpa Sastrawi) maupun Skenario B (Dengan Sastrawi) mencatatkan nilai *Recall* yang sangat terdegradasi pada level 11,6%, yang mengindikasikan terjadinya *False Negative* massal (sistem meloloskan judul yang seharusnya ditolak).

Tabel 3. Hasil Uji Ablasi Pengaruh Prapemrosesan Morfologis (*Threshold* = 0.60)

Skenario Prapemrosesan	Precision	Recall	F1-Score	Keterangan / Analisis Karakteristik
Tanpa Sastrawi (Teks Mentah)	100.0%	11.6%	20.8%	Sistem terlalu kaku; lonjakan <i>False Negative</i> terjadi karena variasi imbuhan dianggap sebagai entitas vektor yang sama sekali berbeda.
Dengan Sastrawi (Kata Dasar)	100.0%	11.6%	20.8%	Reduksi afiksasi Sastrawi gagal mendongkrak matriks karena judul uji didominasi oleh teknik plagiasi parafrasa ekstrem (penggunaan sinonim).

Angka identik pada kedua skenario ini membuktikan sebuah keterbatasan inheren (*inherent limitation*) dari pemodelan *Term Frequency-Inverse Document Frequency* (TF-IDF) ketika dihadapkan pada dimensi teks ultra-pendek (7–15 kata). Normalisasi afiksasi oleh Sastrawi terbukti tidak memberikan dampak signifikan apabila usulan plagiat mahasiswa tidak sekadar memodifikasi imbuhan, melainkan menggunakan teknik parafrasa ekstrem atau substitusi sinonim (misalnya mengubah kata "Pembuatan" menjadi "Rancang Bangun"). TF-IDF akan tetap memproyeksikannya sebagai ruang dimensi vektor yang terpisah. Temuan empiris ini menjadi justifikasi saintifik yang sangat kuat mengapa arsitektur pemrosesan teks tidak dapat bertumpu pada satu nilai ambang batas statis, melainkan menuntut adanya zonasi kelayakan tiga tingkat (*threshold tuning*) agar celah kelemahan leksikal ini dapat diakomodasi oleh intuisi evaluasi manusia.

3.9 Analisis Distribusi Ekstraksi Fitur Leksikal

Kinerja pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) sangat bergantung pada distribusi kosakata (*vocabulary*) di dalam korpus referensi.

Tabel 4. Distribusi 10 Akar Kata dengan Bobot TF-IDF Tertinggi

Peringkat	Akar Kata (Stemmed)	Bobot TF-IDF Agregat	Interpretasi Domain
1	bas	56.05	Akar kata dari "basis" (berbasis). Mengindikasikan penegasan platform atau teknologi dasar yang menopang luaran PkM (misal: berbasis web, berbasis IoT).
2	web	53.50	Platform implementasi teknologi atau luaran perangkat lunak yang paling mendominasi pilihan mahasiswa Teknik Informatika.
3	sistem	49.82	Menggambarkan fokus utama luaran pengabdian berupa perancangan atau rekayasa perangkat lunak.
4	informasi	47.34	Berpasangan erat dengan kata "sistem" (Sistem Informasi), merujuk pada jenis atau spesifikasi produk teknologi utama yang dibangun.
5	tingkat	41.45	Akar kata dari "meningkatkan" atau "peningkatan". Menunjukkan tujuan utama PkM yang berorientasi pada pemberian nilai tambah bagi mitra sasaran.
6	kembang	36.43	Akar kata dari "pengembangan" atau "mengembangkan". Mengindikasikan metode pelaksanaan berupa pembuatan inovasi/sistem baru.
7	smk	35.02	Menunjukkan profil institusi mitra sasaran atau lokus pengabdian yang sangat spesifik dan paling diminati oleh mahasiswa.
8	siswa	34.64	Mewakili entitas pengguna (end-user) atau kelompok masyarakat sasaran yang menerima intervensi teknologi.
9	website	33.03	Merupakan variasi leksikal/sinonim dari "web", menegaskan tingginya tren pembuatan portal sebagai solusi digitalisasi masyarakat.
10	digital	31.75	Menggambarkan tema besar inovasi yang ditawarkan, yakni seputar digitalisasi tata kelola, layanan, atau literasi teknologi mitra.

Untuk memahami karakteristik leksikal dari 839 data primer usulan PkM mahasiswa Teknik Informatika, dilakukan ekstraksi peringkat kata kunci (*Top-N Terms*) berdasarkan akumulasi nilai bobot TF-IDF tertinggi di seluruh dokumen. Tabel Y merangkum 10 akar kata dengan bobot kemunculan paling dominan setelah melalui proses penyaringan *stopword* dan *stemming*.

Berdasarkan Tabel 4, terlihat adanya anomali di mana kata hubung umum bahasa Indonesia seperti "*dan*", "*untuk*", dan "*di*" menduduki peringkat teratas, berdampingan dengan kata yang belum tereduksi afiksasinya seperti "*meningkatkan*" dan "*pengembangan*". Fenomena ini menunjukkan kelemahan fundamental dari pembobotan TF-IDF apabila proses prapemrosesan (khususnya *stopword removal* dan *stemming* Sastrawi) mengalami kebocoran (*bypass*) atau tidak dieksekusi secara ketat pada baris data tertentu. Tingginya frekuensi absolut dari kata tugas ini otomatis membengkakkan matriks korpus dan mengalahkan bobot istilah teknis yang krusial. Temuan ini menjadi justifikasi empiris yang sangat kuat mengapa penelitian ini mewajibkan injeksi modul Sastrawi secara *hardcode* di dalam arsitektur *engine* pencocokan, agar anomali leksikal semacam ini tereliminasi sebelum algoritma *Cosine Similarity* menghitung sudut vektor antar-judul.

3.10 Pengujian Fungsionalitas Operasional (*Black Box Testing*)

Untuk memverifikasi keandalan operasional model di tingkat pengguna akhir, pengujian fungsional berbasis *black box* dilakukan pada antarmuka web Flask. Pengujian dilakukan melalui simulasi ekstrem yang merepresentasikan dinamika pengajuan proposal mahasiswa.



Gambar 7. Respons Antarmuka Web Flask pada Skenario Judul Jenuh (Kiri) dan Judul Unik (Kanan)

Pengujian *Black Box* dirancang untuk memvalidasi presisi dan daya tanggap (*responsiveness*) algoritma ketika diintegrasikan secara utuh ke dalam antarmuka pengguna akhir (*front-end*). Pengujian dieksekusi melalui serangkaian simulasi ekstrem (*boundary value analysis*) untuk mengamati berbagai kemungkinan usulan judul mahasiswa di lapangan.

Skenario 1: Simulasi Usulan Redundan (Kondisi Jenuh)

Pada skenario pertama, dimasukkan usulan judul mahasiswa yang identik secara metode dan objek dengan korpus referensi yang sudah ada di basis data (misalnya: *Peningkatan Efisiensi Administrasi Sekolah melalui Sistem Pendaftaran dan Absensi Digital*). Begitu dikirimkan, kerangka kerja Flask langsung merutekan kueri ke modul pemrosesan (*engine*). Dengan latensi di bawah 1 detik, sistem merender peringatan visual berwarna merah (Sangat Mirip/Jenuh) dengan tingkat persentase kemiripan 100,0%. Simulasi ini membuktikan bahwa algoritma *Cosine Similarity* secara absolut mendeteksi duplikasi total, sehingga efektif memblokir upaya plagiarisme langsung.

Skenario 2: Simulasi Inovasi Baru (Kondisi Novelty Tinggi)

Pada skenario kedua, diinputkan usulan judul yang mengandung entitas teknologi mutakhir yang belum mendominasi pangkalan referensi (misalnya: *Deteksi Kematangan Mangga berbasis IoT*). Sistem dengan seketika merender peringatan visual berwarna hijau dengan tingkat kemiripan sangat rendah, yakni 16,9% (Unik/Layak). Skor yang sangat kecil ini mengonfirmasi kepekaan pembobotan *Inverse Document Frequency* (IDF). Istilah langka seperti "IoT" dan "Mangga" diberikan prioritas bobot spesifik yang tinggi, sehingga membedakannya secara mutlak dari judul-judul lain yang hanya memiliki kesamaan kata hubung atau *stopword* biasa.

Secara keseluruhan, analisis komparasi pengujian fungsional pada kedua skenario ini membuktikan bahwa arsitektur modular yang memisahkan pengendali web (*app.py*), modul prapemrosesan (*preprocessor.py*), dan algoritma klasifikasi (*engine*) telah tersinkronisasi secara ideal. Model sistem seleksi ini terbukti tidak hanya andal sebagai instrumen pencegah plagiaris, melainkan juga adaptif bertindak sebagai asisten pengambil keputusan.

Prototipe secara efektif mengeliminasi subjektivitas penilaian manual dan memfasilitasi percepatan verifikasi orisinalitas luaran inovasi PkM secara *real-time*.

4. KESIMPULAN

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 eksperimen 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.

REFERENSI

- [1] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. draft. Pearson Education.
- [2] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.
- [3] 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, doi: 10.22219/kinetik.v7i4.1405.
- [4] 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, doi: 10.21456/vol13iss1pp45-53.
- [5] 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 (JTIK)*, vol. 9, no. 3, pp. 501–510, doi: 10.25126/jtiik.2022934051.
- [6] 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, doi: 10.33395/sinkron.v8i3.13500.
- [7] 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, doi: 10.46764/teknimedia.v2i2.51.
- [8] 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, doi: 10.29207/resti.v7i1.3501.
- [9] 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, doi: 10.26418/jp.v9i2.50123.



- [10] N. V. A. E. Putri and N. M. A. W. Dewi, "PDF Plagiarism Detection System Using Cosine Similarity," *Jurnal Sains, Sistem, dan Teknologi Komputer (JSSTK)*, vol. 3, no. 1, pp. 1–10, doi: 10.24912/jsstk.v3i1.37103.
- [11] 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, doi: 10.47065/bits.v4i3.2511.
- [12] 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, doi: 10.31000/jika.v9i2.13247.
- [13] 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, doi: 10.24246/aiti.v23i2.319-333.
- [14] 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, doi: 10.30864/eksplora.v12i1.1074.
- [15] 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, doi: 10.22219/repositor.v2i6.30715.
- [16] 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, doi: 10.37034/jsisfotek.v5i3.220.
- [17] 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, doi: 10.46576/device.v6i2.6724.
- [18] S. Pawestri and 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, doi: 10.30865/mib.v8i1.7109.
- [19] 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, doi: 10.22146/jnteti.v12i2.4502.
- [20] 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, doi: 10.47709/brilliance.v4i2.5547.
- [21] 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, doi: 10.24002/jbi.v15i1.7501.
- [22] D. Iskandar and A. Kurniawati, "Analisis Perbandingan Teknik Word2vec dan Doc2vec dalam Mengukur Kemiripan Dokumen Menggunakan Cosine Similarity," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 1, pp. 133–144, doi: 10.25126/jtiik.20251219143.
- [23] 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, doi: 10.21609/jiki.v16i2.1105.
- [24] S. A. Gilbert and M. I. Sulisty, "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, doi: 10.55123/storage.v2i3.2333.
- [25] S. A. S. Mola, Y. C. Luttu, and D. N. Rumlaklak, "Perbandingan Metode Machine Learning dalam Analisis Sentimen Komentar Pengguna Aplikasi InDriver pada Dataset Tidak Seimbang," *Jurnal Sistem Informasi Bisnis*, vol. 14, no. 3, pp. 247–255, doi: 10.21456/vol14iss3pp247-255.
- [26] D. L. Girsang, A. Sidiq, and T. S. Elenaputri, "Analisis Sentimen Masyarakat terhadap Layanan BPJS Kesehatan dan Faktor-Faktor Pendukung Opini dengan Pemodelan Natural Language Processing (NLP)," *Emerging Statistics and Data Science Journal*, vol. 1, no. 2, pp. 238–249, doi: 10.20885/esds.vol1.iss.2.art24.