(2025), 3 (6): 141-160 **Scientica** 3021-8209

Jurnal Ilmiah Sain dan Teknologi

IDENTIFIKASI DOKUMEN TUGAS AKHIR BERDASARKAN KELOMPOK KEAHLIAN MENGGUNAKAN *TERM FREQUENCY* DAN *KEYWORD MATCHING*

Dhimas Rangga Irza Ruizsandi [#1], Arif Bijaksana Putra Negara [#2], Yus Sholva [#3]

Jurusan Informatika, Fakultas Teknik, Universitas Tanjungpura

Jl. Prof. Dr. H. Hadari Nawawi, Pontianak

[1]dhimasruiz@gmail.com

Abstrak (Indonesia)

Peningkatan jumlah dokumen tugas akhir setiap tahunnya menuntut adanya sistem yang mampu mengelola dan mengidentifikasi dokumen secara otomatis, khususnya dalam mengidentifikasi berdasarkan kelompok keahlian. Penelitian ini bertujuan untuk membangun sebuah sistem identifikasi dokumen tugas akhir berbasis web dengan mengimplementasikan metode *Term Frequency (*TF) dan pencocokan kata kunci (*keyword matching*). Dokumen diolah melalui beberapa tahap, yaitu parsing bagian penting dokumen (judul, abstrak, latar belakang, dan kesimpulan), *preprocessing* teks, vektorisasi TF, serta pencocokan terhadap daftar kata kunci tetap dan kata kunci dinamis. Bobot TF dikalikan dengan bobot kecocokan berdasarkan jenis pencocokan, yaitu *exact match*, *partial match*, dan *title match*, untuk menghasilkan skor identifikasi setiap kelompok keahlian. Selain itu, sistem juga dilengkapi fitur pencarian dokumen serupa menggunakan *cosine similarity* dan mekanisme pembaruan kata kunci dinamis berdasarkan dokumen yang telah teridentifikasi. Evaluasi sistem dilakukan menggunakan *confusion matrix* yang menunjukkan tingkat akurasi sebesar 89,5% serta nilai *precision* yang tinggi di seluruh kelas. Hasil penelitian ini menunjukkan bahwa pendekatan yang digunakan efektif dalam mengidentifikasi dan mengelompokkan dokumen tugas akhir.

Sejarah Artikel

Submitted: 11 September 2025

Accepted: 14 September 2025

Published: 15 September 2025

Kata Kunci

identifikasi dokumen, term frequency, *keyword matching*, *cosine similarity*, sistem informasi.

I. PENDAHULUAN

Perkembangan ilmu pengetahuan dan teknologi yang pesat berdampak pada meningkatnya jumlah dokumen tugas akhir di perguruan tinggi setiap tahunnya. Di Universitas Tanjungpura, khususnya pada Program Studi Sarjana Informatika, puluhan mahasiswa menyelesaikan skripsi yang mencakup berbagai topik dalam bidang teknologi informasi. Kondisi ini menimbulkan tantangan baru dalam pengelolaan, pengarsipan, dan pengelompokan dokumen secara sistematis dan efisien. Proses identifikasi kelompok keahlian dokumen tugas akhir yang selama ini dilakukan secara manual oleh mahasiswa dan dosen pembimbing berpotensi menghasilkan ketidakkonsistenan karena bergantung pada penilaian subjektif tanpa analisis konten yang terstruktur.

Tugas akhir merupakan dokumen ilmiah yang mencerminkan hasil penelitian mahasiswa sesuai bidang keahliannya [1]. Pada Program Studi Sarjana Informatika Universitas Tanjungpura, bidang keahlian dibagi menjadi empat rumpun utama: *Computation & Artificial Intelligence*, *Networking & Security*, *Software Engineering & Mobile Computing*, dan *Information Systems & Data Spatial*. Ketidaktepatan dalam mengidentifikasi kelompok keahlian dapat berdampak pada kesesuaian referensi serta pemanfaatan dokumen sebagai sumber informasi akademik. Tanpa sistem identifikasi berbasis konten, informasi penting dalam skripsi sulit diakses secara terstruktur, sehingga mengurangi nilai guna dokumen sebagai *repository* ilmiah yang potensial untuk analisis tren keilmuan maupun penelitian lanjutan.

Untuk mengatasi permasalahan ini, penelitian ini mengusulkan pengembangan sistem identifikasi dokumen tugas akhir berbasis *Term Frequency* (TF) dan *keyword matching*.

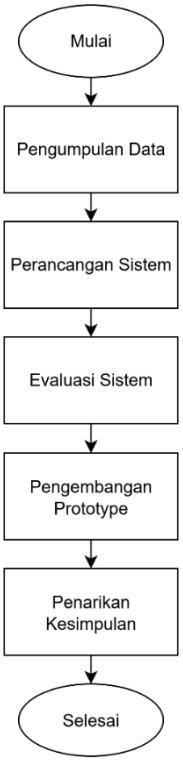
Metode TF digunakan untuk memberikan bobot terhadap kata-kata penting dalam dokumen [2], sedangkan *keyword matching* memanfaatkan daftar kata kunci tetap yang divalidasi oleh dosen dari masing-masing kelompok keahlian [3]. Kombinasi ini diharapkan mampu meningkatkan akurasi identifikasi, khususnya pada kasus tumpang tindih istilah antar kelompok keahlian. Sistem juga dilengkapi fitur pencarian dokumen serupa menggunakan *cosine similarity* [4], yang menghitung kedekatan konten antar dokumen sehingga dapat memvalidasi hasil identifikasi sekaligus membantu pengguna menemukan referensi relevan.

Beberapa penelitian sebelumnya menunjukkan efektivitas pendekatan ini. Penelitian oleh [5] memanfaatkan TF-IDF dalam aplikasi *repository* skripsi untuk meningkatkan pencarian dokumen. Penelitian yang dilakukan oleh [6] menerapkan kombinasi TF-IDF dan *cosine similarity* untuk mengelompokkan judul seminar proposal dengan hasil yang akurat. Dan [7] juga membuktikan bahwa TF-IDF mampu mengelompokkan teks dalam dataset tidak terstruktur dengan kinerja yang baik.

Tujuan dari penelitian ini adalah merancang dan mengimplementasikan sistem identifikasi dokumen tugas akhir yang mampu mengelompokkan dokumen secara tepat sesuai bidang keahlian yang telah ditentukan. Sistem ini memanfaatkan pembobotan kata menggunakan metode *Term Frequency* (TF) untuk menonjolkan istilah penting, serta teknik pencocokan kata kunci tetap yang telah diverifikasi oleh dosen. Selain itu, sistem dilengkapi mekanisme kata kunci dinamis yang dihasilkan secara otomatis dari dokumen untuk memperluas cakupan pencocokan istilah. Penelitian ini juga bertujuan menyediakan fitur pencarian dokumen serupa berbasis *cosine similarity*, sehingga dapat mendukung validasi hasil identifikasi sekaligus membantu pengguna menemukan referensi yang relevan.

II. METODOLOGI PENELITIAN

Penelitian ini terdiri dari serangkaian proses yang berurutan, yaitu pengumpulan data, perancangan sistem, evaluasi sistem, pengembangan *prototype* dan penarikan kesimpulan. Setiap proses meliputi metode dan pendekatan yang lebih mendalam untuk kebutuhan proses berikutnya. Adapun alur penelitian diilustrasikan pada Gambar 1.



Gambar 1. Alur penelitian

A. *Pengumpulan Data*

Data dalam penelitian ini diperoleh dari berbagai sumber. Dokumen tugas akhir dalam format DOCX dan PDF dikumpulkan melalui *Google Drive* milik Program Studi Sarjana Informatika Universitas Tanjungpura. Dokumen-dokumen tersebut merupakan arsip administrasi yang diserahkan oleh mahasiswa sebagai syarat mengikuti sidang akhir. Selain itu, digunakan pula template skripsi tahun 2015 Fakultas Teknik yang diunduh dari laman Unduhan Sistem Pendukung Outline Tugas Akhir (SPOTA). Template ini dijadikan acuan dalam perancangan regex pada proses *document parsing*. Sementara itu, kumpulan kata kunci tetap (*fixed keywords*) yang merepresentasikan masing-masing kelompok keahlian ditentukan melalui observasi dan analisis, serta divalidasi oleh dosen ahli dari setiap kelompok keahlian yang bersangkutan.

B. *Perancangan Sistem*

Tahapan perancangan sistem dilakukan untuk merancang struktur dan alur kerja dari sistem yang akan dibangun. Pada penelitian ini, sistem dirancang untuk menjalankan tiga sistem yang akan diimplementasikan pada aplikasi yaitu identifikasi kelompok keahlian dengan *keyword matching*, pencarian dokumen serupa dengan *cosine similarity*, dan kata kunci dinamis.

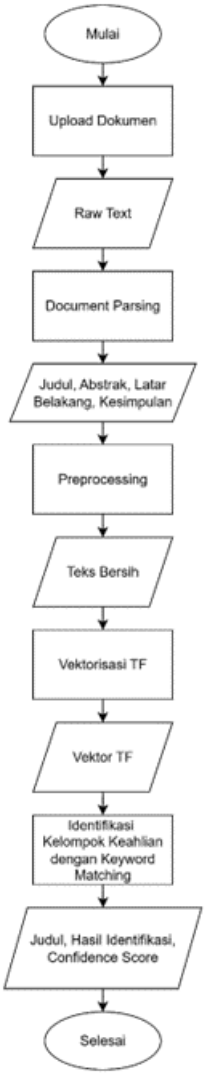
1) Identifikasi Dokumen dengan Keyword Matching

Tahapan ini merupakan inti penelitian yang berfokus pada pengembangan sistem identifikasi dokumen tugas akhir berdasarkan kelompok keahlian. Berdasarkan ilustrasi pada Gambar 2, proses dimulai dari *document parsing* untuk mengekstrak bagian penting dokumen, yaitu judul, abstrak, latar belakang, dan kesimpulan menggunakan *regular expression* (regex) [8]. Bagian tersebut dipilih karena mengandung inti permasalahan, tujuan penelitian, dan kesimpulan yang relevan untuk penentuan topik. Regex disesuaikan dengan variasi format dokumen pada Template Skripsi Fakultas Teknik Universitas Tanjungpura dan dokumen aktual, sehingga sistem tetap fleksibel meskipun format tidak seragam.

Hasil parsing kemudian diproses melalui *preprocessing* [9] yang meliputi pembersihan teks [10], penghapusan *stopword*, *tokenization* [11], dan *lemmatization* [12] untuk menghasilkan teks bersih yang siap diolah. Seluruh bagian yang telah diproses digabungkan menjadi satu string teks dan dikonversi menjadi vektor numerik menggunakan metode TF. Representasi vektor ini menjadi dasar untuk proses *keyword matching* dalam penentuan kelompok keahlian.

Pada tahap identifikasi, sistem mencocokkan kata-kata dalam vektor dokumen dengan daftar kata kunci tetap dan dinamis. Kata kunci tetap dianalisis dan divalidasi oleh dosen dari setiap kelompok keahlian, sedangkan kata kunci dinamis diperoleh dari analisis token dengan bobot TF-IDF tinggi pada dokumen sebelumnya yang telah teridentifikasi. Proses pencocokan mempertimbangkan tiga jenis kecocokan [13][14] yaitu, *exact match* yang merupakan kata yang sama persis dengan kata kunci, bobot TF dikalikan 3. *Partial match* dimana kata memiliki bagian dari substring sebuah kata, bobot TF dikalikan 1,5. Dan *title match* yang dimana kata kunci ditemukan di judul, bobot akhir ditingkatkan sebesar 1 + 0,3 x n, dimana n adalah jumlah kata kunci yang cocok pada judul.

Skor akhir untuk setiap kelompok keahlian dihitung dengan menjumlahkan bobot token yang sesuai, memperhitungkan nilai TF, jenis kecocokan, dan kontribusi judul. Kelompok dengan skor tertinggi ditetapkan sebagai hasil identifikasi.



Gambar 2. Alur proses identifikasi kelompok keahlian

2) Pencarian Dokumen Serupa dengan Cosine Similarity

Tahapan ini dilakukan setelah sistem berhasil mengidentifikasi kelompok keahlian dari dokumen tugas akhir input. Setelah diketahui kelompok keahliannya, sistem akan melakukan pencarian dokumen serupa hanya pada koleksi dokumen yang telah teridentifikasi dalam kelompok keahlian yang sama. Pendekatan ini bertujuan untuk menjaga relevansi hasil pencarian dan mempersempit ruang pencocokan agar lebih kontekstual terhadap bidang keilmuan yang sesuai.

Gambar 3. Alur proses pencarian dokumen serupa dengan **cosine similarity**

Setiap dokumen baik dokumen input maupun dokumen dalam database akan divektorisasi dalam bentuk vektor TF-IDF. Dengan memanfaatkan vektor ini, sistem dapat mengukur tingkat kemiripan antar dokumen berdasarkan distribusi kata-kata penting dalam ruang vektor multidimensi. Sesuai pada Gambar 3. dapat dilihat tahapan dimulai dengan representasi dokumen input dalam bentuk matriks TF-IDF, di mana setiap kata dalam dokumen diubah menjadi vektor numerik berdasarkan frekuensi dan bobot informasinya.

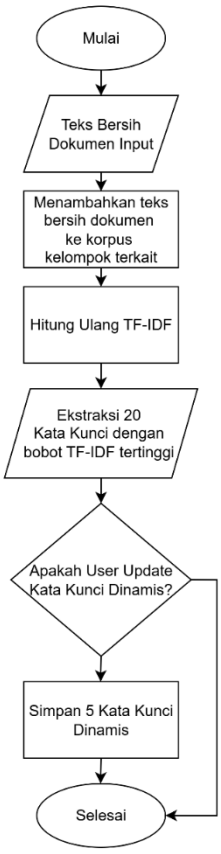
Selanjutnya, sistem akan melakukan proses pengambilan koleksi dokumen dari database yang termasuk dalam kelompok keahlian yang sama dengan dokumen input. Proses ini penting untuk memastikan bahwa pencarian dilakukan dalam ruang lingkup yang relevan dan terarah. Setelah koleksi dokumen berhasil diambil, dilakukan proses perhitungan *cosine similarity* antara vektor dokumen input dengan vektor dari setiap dokumen dalam database.

Hasil dari proses ini berupa skor similaritas untuk setiap dokumen dalam kelompok keahlian. Sebagai output, sistem menghasilkan daftar dokumen serupa beserta skor kemiripannya, dokumen-dokumen dengan nilai kemiripan tertinggi dianggap paling relevan dan dapat digunakan sebagai referensi atau bahan perbandingan oleh pengguna.

3) Kata Kunci Dinamis

Sistem identifikasi dokumen dalam penelitian ini dilengkapi dengan fitur kata kunci dinamis (*dynamic keywords*) sebagai pelengkap dari kata kunci tetap (*fixed keywords*). Tidak seperti kata kunci tetap yang ditentukan secara manual melalui observasi dan validasi oleh dosen dari masing-masing kelompok keahlian, kata kunci dinamis bersifat adaptif dan diperoleh secara otomatis melalui analisis isi dokumen yang telah berhasil diidentifikasi oleh sistem. Dengan demikian, fitur ini memungkinkan sistem untuk mengenali istilah-istilah baru atau *domain-specific* yang mungkin belum tercakup dalam kata kunci tetap, serta menyesuaikan sistem terhadap dinamika istilah dalam dokumen akademik.

Setelah dokumen berhasil diidentifikasi ke dalam salah satu kelompok keahlian, sistem akan menjalankan proses pembaruan kata kunci dinamis sebagaimana ditunjukkan pada Gambar 4. Dibawah. Proses ini diawali dengan penggunaan teks dokumen yang telah melalui tahap *preprocessing*, Teks bersih ini kemudian ditambahkan ke dalam korpus dokumen kelompok keahlian yang telah teridentifikasi sebelumnya, sehingga dapat memengaruhi nilai IDF dalam ruang lingkup kelompok tersebut.



Gambar 4. Alur proses memperbarui kata kunci dinamis

Selanjutnya, sistem menghitung ulang nilai TF-IDF untuk seluruh dokumen dalam kelompok keahlian, termasuk dokumen baru yang ditambahkan. Berdasarkan hasil vektorisasi tersebut, sistem mengekstrak 20 kata dengan bobot TF-IDF tertinggi dari korpus kelompok keahlian tersebut sebagai kandidat kata kunci dinamis. Kemudian, sistem akan menunggu respon pengguna untuk memperbarui atau tidak kata kunci tersebut. Jika pengguna memilih untuk memperbarui, maka sistem akan menyimpan 5 kata kunci terbaik ke dalam basis data.

Proses ini bersifat kondisional, dimana jika belum terdapat dokumen yang berhasil diidentifikasikan ke dalam suatu kelompok keahlian, maka sistem tidak akan menghasilkan kata kunci dinamis untuk kelompok tersebut. Dengan demikian, integrasi fitur ini tidak hanya memperkaya proses identifikasi dokumen, tetapi juga meningkatkan fleksibilitas dan akurasi sistem dalam mengenali konteks tematik, terutama pada dokumen yang menggunakan variasi istilah yang tidak selalu dapat dijangkau oleh kata kunci tetap. Selain itu, hasil dari kata kunci dinamis ini juga disimpan secara terstruktur sebagai data pendukung yang akan digunakan dalam iterasi proses identifikasi berikutnya.

C. *Evaluasi Sistem*

Pada tahap ini, sistem akan divalidasi dengan tujuan untuk mengukur sejauh mana performa sistem dalam mengidentifikasi dokumen secara akurat ke dalam kelompok keahlian yang tepat. Proses evaluasi ini dilakukan dengan membandingkan hasil prediksi sistem terhadap label kebenaran (*ground truth*) yang telah ditetapkan sebelumnya. Dalam penelitian ini, evaluasi dilakukan menggunakan *confusion matrix*, yaitu suatu tabel yang menunjukkan jumlah prediksi benar dan salah yang dilakukan pada masing-masing kelas.

TABEL I
METRIK EVALUASI

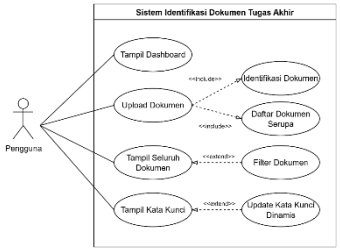
Label	Precision
Computation & Artificial Intelligence	
Information System & Data Spatial	
Networking & Security	
Software Eng. & Mobile Computing	
Accuracy	

Tabel 1. diatas menunjukkan metrik evaluasi yang digunakan untuk menilai performa sistem dalam melakukan identifikasi dokumen, yaitu *accuracy*. *Accuracy* merupakan metrik yang mengukur proporsi hasil identifikasi yang benar terhadap seluruh data uji yang tersedia. Metrik ini digunakan untuk memberikan gambaran umum mengenai sejauh mana sistem mampu mengenali kelompok keahlian dokumen secara tepat. Meskipun *accuracy* memiliki keterbatasan, metrik ini tetap digunakan dalam penelitian ini sebagai indikator awal untuk mengukur kinerja sistem secara keseluruhan. Hasil *accuracy* dapat memberikan informasi dasar mengenai efektivitas sistem identifikasi sebelum dilakukan evaluasi lanjutan yang lebih spesifik.

D. *Pengembangan Prototype*

Tahap ini merupakan proses implementasi sistem, di mana seluruh komponen metode yang telah dirancang diintegrasikan ke dalam sebuah aplikasi berbasis web. Pengembangan dilakukan dengan memanfaatkan *framework* Django [15] berbasis bahasa pemrograman Python [16], yang mendukung manajemen data, proses sistem, serta pengembangan antarmuka pengguna secara terstruktur.

1) Perancangan Use Case



***Gambar 5.* Use Case Diagram**

Berdasarkan **Gambar 3.12**, ditunjukkan bahwa pengguna memiliki akses terhadap delapan *use case* utama, antara lain: menampilkan dashboard, mengunggah dokumen, melakukan identifikasi dokumen dengan *keyword matching*, menampilkan daftar dokumen serupa berdasarkan *cosine similarity*, melihat seluruh dokumen yang telah terproses, memfilter dokumen berdasarkan kategori tertentu, menampilkan kata kunci (baik tetap maupun dinamis), serta melakukan pembaruan kata kunci dinamis berdasarkan hasil pembobotan TF-IDF.

Penjelasan lebih detail mengenai *use case* dapat dilihat pada Tabel 2. dibawah.

TABEL II
KETERANGAN USE CASE DIAGRAM

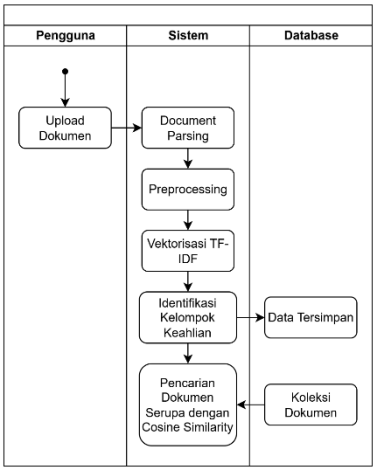
No	Use Case	Keterangan
1	Tampil Dashboard	Menampilkan tampilan awal sistem yang berisi ringkasan dan menu navigasi utama.
2	Upload Dokumen	Mengunggah dokumen tugas akhir dalam format PDF atau DOCX ke dalam sistem.
3	Identifikasi Dokumen	Melakukan proses identifikasi kelompok keahlian berdasarkan keyword matching.
4	Daftar Dokumen Serupa	Menampilkan dokumen-dokumen serupa berdasarkan hasil perhitungan cosine similarity.
5	Tampil Seluruh Dokumen	Menampilkan daftar seluruh dokumen yang telah diproses oleh sistem.
6	Filter Dokumen	Menyaring dokumen berdasarkan kelompok keahlian atau kriteria tertentu.
7	Tampil Kata Kunci	Menampilkan daftar kata kunci tetap dan dinamis untuk masing-masing kelompok.
8	Update Kata Kunci Dinamis	Memperbarui kata kunci dinamis berdasarkan hasil pembobotan TF-IDF terbaru.

2) Perancangan Activity Diagram

Activity diagram bertujuan untuk menggambarkan alur aktivitas sistem secara dinamis, mulai dari input pengguna hingga sistem menghasilkan output. Diagram ini memvisualisasikan urutan proses dan keputusan logis yang terjadi saat pengguna menjalankan suatu fungsi dalam sistem.

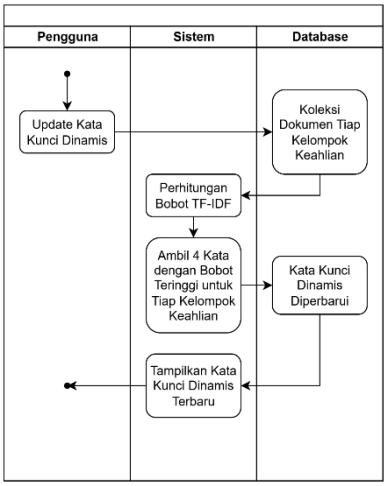
Gambar 6. dibawah menggambarkan alur proses sistem identifikasi dokumen tugas akhir mulai dari saat pengguna mengunggah dokumen hingga sistem menampilkan dokumen serupa. Proses diawali dari pengguna yang mengunggah dokumen ke sistem. Proses diawali dari pengguna yang mengunggah dokumen ke sistem. Sistem kemudian melakukan parsing dokumen untuk mengekstrak bagian, dilanjutkan dengan *preprocessing* teks. Hasil *preprocessing* kemudian diubah menjadi vektor TF. Selanjutnya, sistem mengidentifikasi

kelompok keahlian dokumen menggunakan *keyword matching*, dan hasil identifikasi disimpan ke database. Terakhir, sistem melakukan pencarian dokumen serupa berdasarkan hasil identifikasi kelompok keahlian dengan metode *cosine similarity* terhadap koleksi dokumen yang relevan dalam database.



Gambar 6.* Activity Diagram *identifkasi dokumen

Selanjutnya, pada Gambar 7. dibawah dapat dilihat setelah sistem berhasil mengidentifikasi dan mengelompokkan dokumen ke dalam kelompok keahlian tertentu, seluruh dokumen yang tergolong dalam kelompok tersebut akan diolah ulang menggunakan metode TF-IDF untuk menghitung bobot kemunculan setiap term. Dari hasil pembobotan tersebut, sistem akan memilih beberapa kata dengan nilai TF-IDF tertinggi yang dianggap paling representatif terhadap topik atau domain dari kelompok tersebut. Kata-kata ini kemudian digunakan untuk memperbarui daftar kata kunci dinamis pada kelompok keahlian yang bersangkutan.

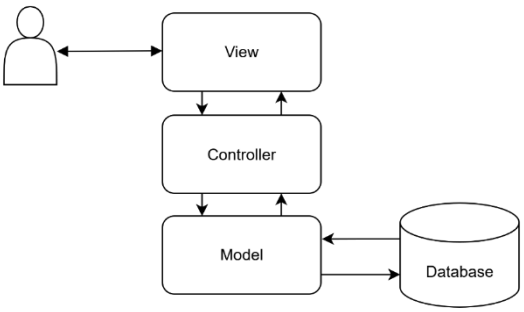


Gambar 7* Activity Diagram *perbarui kata kunci dinamis

3) Perancangan Arsitektur Sistem

Perancangan arsitektur sistem dilakukan dengan mengadopsi pendekatan Model-View-Controller (MVC) guna memisahkan tanggung jawab antar komponen, sehingga pengembangan lebih terstruktur dan mudah dipelihara. Komponen *Model* menangani logika pemrosesan data seperti *preprocessing*, vektorisasi TF, identifikasi kelompok keahlian, dan pencarian dokumen serupa. Komponen *View* bertanggung jawab menyajikan antarmuka pengguna seperti halaman unggah dokumen, hasil identifikasi, dan daftar kata kunci. Sementara itu, *Controller* mengatur alur interaksi antara *Model* dan *View*. Gambar 8. dibawah

yang merupakan arsitektur sistem mendukung modularitas dan skalabilitas sistem, serta memungkinkan pengembangan fitur tambahan secara fleksibel.



Gambar 8. Arsitektur system

4) Implementasi system

Tahapan ini akan mengintegrasikan kode Python ke dalam komponen *Model* dalam arsitektur MVC, yang diimplementasikan menggunakan *framework* Django. Komponen *Model* bertanggung jawab terhadap logika bisnis sistem, termasuk seluruh proses analisis teks, perhitungan, dan penyimpanan data. Dalam konteks penelitian ini, kode Python yang digunakan untuk proses seperti *document parsing* dengan regex, *text preprocessing*, vektorisasi TF, *keyword matching*, dan perhitungan *cosine similarity*, seluruhnya dimasukkan dalam modul Python (di folder utils/ atau services/). Fungsi-fungsi ini kemudian dipanggil oleh komponen *Controller* (views.py) saat pengguna mengunggah dokumen melalui antarmuka *View* (template HTML).

5) Blackbox testing

Pengujian sistem dilakukan menggunakan *blackbox testing*, di mana setiap fungsionalitas sistem diuji berdasarkan input dan output yang diharapkan tanpa melihat struktur kode internal. Berikut ringkasan metode pengujian berdasarkan Tabel 3. dibawah.

Pengujian difokuskan pada fitur-fitur utama sistem sesuai dengan *use case* pada Gambar 5. untuk memastikan bahwa sistem berfungsi sesuai dengan spesifikasi yang ditetapkan. Beberapa skenario pengujian dilakukan untuk mencakup seluruh alur penggunaan aplikasi, mulai dari akses awal hingga proses identifikasi dokumen dan manajemen kata kunci

TABEL III
BLACKBOX TESTING

No	Pengujian	Input Uji	Ekspektasi Output	Status
1	Tampil Dashboard	Pengguna membuka halaman utama	Sistem menampilkan ringkasan statistik dan menu navigasi utama	
2	Upload Dokumen	File .pdf atau .docx valid	Dokumen berhasil diunggah dan ditampilkan untuk identifikasi	
3	Upload Dokumen (invalid)	File .jpg, atau .xlsx	Sistem menolak file dan menampilkan pesan kesalahan	
4	Identifikasi Dokumen	Dokumen hasil unggah	Sistem memproses dan menampilkan judu, hasil identifikasi, confidence score dan daftar dokumen yang mirip	
5	Tampil Seluruh Dokumen	Klik menu "Daftar Dokumen"	Sistem menampilkan seluruh dokumen yang sudah diidentifikasi	

6	Filter Dokumen	Pilih filter berdasarkan kelompok keahlian	Sistem menampilkan dokumen sesuai dengan filter yang dipilih	
7	Delete dokumen	Klik tombol delete pada suatu dokumen	Sistem menghapus dokumen	
8	Tampil Kata Kunci	Klik menu "Daftar Kata Kunci"	Sistem menampilkan daftar kata kunci tetap dan dinamis	
9	Update Kata Kunci Dinamis	Tersedia cukup dokumen untuk setiap kelompok	Sistem menghasilkan kata kunci dinamis berbasis TF-IDF	
10	Update Kata Kunci Dinamis	Belum ada dokumen untuk kelompok tertentu	Sistem menampilkan notifikasi atau tidak melakukan update	
11	Edit Kata Kunci Tetap	Kata kunci tetap sebanyak tepat 15 kata kunci	Sistem memperbarui kata kunci tetap sesuai input	
12	Edit Kata Kunci Tetap	Kata kunci tetap tidak tepat sebanyak 15 kata kunci	Sistem menampilkan notifikasi kata kunci harus 15 kata kunci	

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data dalam penelitian ini dikumpulkan dari tiga sumber. Pertama, dokumen tugas akhir mahasiswa dikumpulkan dari *Google Drive* milik Program Studi Sarjana Informatika Universitas Tanjungpura. Dokumen-dokumen ini merupakan syarat administrasi sidang akhir, berjumlah 87 file dalam format PDF dan DOCX dengan masing-masing pembagian yang dapat dilihat pada Tabel 4**.**

TABEL IV
HASIL PENGUMPULAN DOKUMEN TUGAS AKHIR

No.	Kelompok Keahlian	Jumlah Dokumen
1	Information System & Data Spatial	39
2	Computation & Artificial Intelligence	30
3	Software Engineering & Mobile Computing	16
4	Networking & Security	2
	Total	87

Kedua, template skripsi tahun 2015 dari Fakultas Teknik diunduh melalui laman resmi SPOTA (Sistem Pendukung Outline Tugas Akhir). Template ini berfungsi sebagai acuan utama dalam merancang regex pada Tabel 5. yang digunakan pada tahap *document parsing*. Dengan adanya struktur baku pada template tersebut, proses identifikasi dan ekstraksi bagian-bagian penting dari dokumen skripsi, yaitu judul, abstrak, latar belakang, dan kesimpulan dapat dilakukan dengan lebih sistematis dan konsisten. Pola yang ditemukan dari template kemudian digunakan untuk menyesuaikan sistem terhadap dokumen-dokumen asli yang memiliki format serupa.

TABEL V
PENENTUAN REGEX DARI TEMPLATE SKRIPSI 2015

No	Bagian dari Dokumen	Regex
1	Judul	r'(?im)^(?:[A-Z][A-Z\s]+\n)?([^\n]+?(?=\s*(?:ABSTRAK|ABSTRACT|SKRIPSI|PROGRAM STUDI)))'
2	Abstrak	r'(?im)(?:ABSTRAK|ABSTRACT)[\s.:-]*\s*([^\n].*?)\s*(?=(?:Kata Kunci|Keywords|Latar Belakang|BAB I))'
3	Latar Belakang	r'(?is)(?:Latar\s+Belakang)[\s.:,-]*\s*([\s\S]+?)(?=\s*(?:Perumusan\s+Masalah|Rumusan\s+Masalah))'
4	Kesimpulan	r'(?is)(?:KESIMPULAN DAN SARAN)[\s.:,-]*\s*([\s\S]+?)(?=\s*(?:Saran|Daftar\s+Pustaka))'

Ketiga, kata kunci tetap (*fixed keyword*) merupakan kumpulan kata atau frasa representatif yang disusun dan ditentukan berdasarkan observasi, dan analisis topik dari dokumen yang ada. Proses penentuan kata kunci ini tidak dilakukan secara sembarangan, melainkan melalui diskusi dan justifikasi akademik dengan dosen-dosen untuk memastikan bahwa setiap kata kunci benar-benar mencerminkan karakteristik, cakupan materi, dan fokus riset dari masing-masing kelompok keahlian. Setiap kelompok keahlian memiliki sebanyak 15 kata kunci utama yang dianggap mampu mewakili topik-topik khas dari kelompok tersebut. Kata kunci ini digunakan sebagai dasar dalam proses identifikasi dokumen, pencocokkan antara dokumen dengan kata kunci kelompok keahlian akan menjadi indikator untuk menentukan identifikasi dokumen ke dalam kelompok keahlian yang paling sesuai. Tabel 6. menunjukkan kata kunci tetap dari tiap kelompok keahlian yang telah dikumpulkan sebagai dasar proses identifikasi dokumen.

TABEL VI
HASIL PENGUMPULAN KATA KUNCI TETAP

Kelompok Keahlian	Kata Kunci Yang Dikumpulkan	Kata Kunci Yang Digunakan
Computation & Artificial Intelligence	akurasi, analisis, algoritma, model, deteksi, sensor, rekomendasi, klasifikasi, perbandingan, machine learning, internet of things, prediksi, learning	akurasi, analisis, algoritma, model, deteksi, sensor, rekomendasi, klasifikasi, perbandingan, machine learning, internet of things, prediksi, learning
Networking & Security	cyber, cybersecurity, wireshark, QoS, latency, throughput, DDoS, jaringan, network, firewall, enkripsi, kriptografi, bandwidth, VPN, security, penetration testing, authentication, IDS, IPS, zero trust, TLS, IoT,Injection, Hacking, vulnerability, dork,wireless, sensor, cloud, bruteforce, internet, intranet, client, server, peer to peer, blockchain, trustless system, malware, information security, information governance.	cyber, wireshark, blockchain, latency, throughput, ddos, jaringan, blockchain, network, enkripsi, server, security, bandwidth, kriptografi, aman,

Information System & Data Spatial	peta, location based service, sistem informasi, gis, lokasi, website, informasi, spasial, visualisasi, web, manajemen, tata kelola, sig, manajemen basis data, data spasial, pengambilan keputusan berbasis data, big data analytics, location-based services, smart city, internet of things (iot), data visualization, geospatial intelligence, web gis, data mining, sistem pendukung keputusan (spk), machine learning, integrasi sistem informasi	peta, location based service, sistem informasi, gis, lokasi, website, informasi, spasial, visualisasi, web, manajemen, tata kelola,
Software Engineering & Mobile Computing	ui, pemrograman, android, mobile, ios, kotlin, pengembangan, aplikasi, game, ux, perangkat lunak, agile, deployment, rpl, scrum, desain aplikasi, rancang bangun	ui, pemrograman, android, mobile, ios, kotlin, pengembangan, game, ux, perangkat lunak, agile, deployment, rpl, scrum, rancang bangun

B. Hasil Implementasi Sistem Identifikasi Dokumen

Pada tahap ini, seluruh metode yang telah dirancang dalam penelitian diterapkan dalam bentuk sistem identifikasi dokumen tugas akhir. Fokus utama implementasi adalah mengintegrasikan setiap komponen yang mendukung proses identifikasi kelompok keahlian berdasarkan isi dokumen, mulai dari tahapan *parsing* dokumen, *preprocessing* teks, pembentukan vektor TF, dan pencocokan kata kunci (*keyword matching*).

1) Document Parsing

Langkah pertama dalam proses identifikasi adalah mengekstrak bagian-bagian penting dari dokumen tugas akhir, yaitu: judul, abstrak, latar belakang, dan kesimpulan. Proses ekstraksi dilakukan menggunakan metode regex yang disesuaikan dengan struktur dokumen pada template skripsi 2015.

TABEL VII
CONTOH HASIL PARSING

Bagian Dokumen	Hasil Parsing
Judul	APLIKASI KAWASAN RAWAN BANJIR DAN REKOMENDASI TEMPAT EVAKUASI MENGGUNAKAN DATA DEM DAN BERDASARKAN JARAK TERDEKAT
Abstrak	“Banjir adalah salah satu bencana alam yang disebabkan oleh … sehingga layak digunakan untuk memberikan informasi kawasan banjir.”
Latar Belakang	“Banjir adalah salah satu bencana alam yang disebabkan oleh beberapa faktor alam … menentukan rute evakuasi terdekat dari masyarakat yang terkena bencana banjir di Kabuapaten Sekadau.”
Kesimpulan	“Berdasarkan hasil analisis dan pengujian terhadap perangkat lunak aplikasi … memberikan notifikasi sesuai perintah masing-masing halaman.”

Tabel 7. menampilkan contoh output proses parsing untuk satu dokumen tugas akhir berjudul D1041151049_ANYAP LIUS_LAPORAN TA dalam format dokumen DOCX. Pada tabel tersebut, setiap baris merepresentasikan salah satu atribut dari objek ExtractedSections, yaitu judul, abstrak, latar belakang, dan kesimpulan, yang diekstrak secara otomatis menggunakan

pola regex. Tabel ini menunjukkan bahwa sistem mampu mengenali dan memisahkan keempat bagian utama dari dokumen secara terstruktur, sehingga siap digunakan dalam tahap berikutnya.

2) Preprocessing

Setelah dokumen berhasil di*parsing*, seluruh bagian penting yang diekstraksi yaitu judul, abstrak, latar belakang, dan kesimpulan akan digabung menjadi satu kesatuan teks sebagai input untuk tahap *preprocessing*.

Proses *preprocessing* dilakukan menggunakan fungsi lemmatize dari *library* IndoNLP, yang secara otomatis mencakup beberapa tahapan utama, yaitu *case folding*, penghilangan karakter khusus, dan *lemmatization*. Setelah itu, teks yang telah dibersihkan akan ditokenisasi menggunakan fungsi word_tokenize, lalu setiap token akan dicek dan difilter menggunakan daftar *stopwords* Bahasa Indonesia untuk menghapus kata-kata umum yang tidak membawa makna penting. Akhirnya, daftar token hasil akhir akan digabungkan kembali menjadi satu string teks bersih untuk digunakan pada tahap selanjutnya. Tabel 8. menyajikan perbandingan antara teks sebelum dan sesudah proses *preprocessing* dilakukan.

TABEL VIII
CONTOH HASIL PREPROCESSING

Teks sebelum dipreprocess (824 kata)	Teks setelah dipreprocess (619 kata)
“APLIKASI KAWASAN RAWAN BANJIR DAN REKOMENDASI TEMPAT EVAKUASI MENGGUNAKAN DATA DEM DAN BERDASARKAN JARAK TERDEKAT Banjir adalah salah satu bencana alam yang disebabkan oleh beberapa faktor alam diantaranya dapat berupa curah hujan yang tinggi, kelerengan tanah yang lebih rendah dari permukaan air laut maupun dikarenakan tanggul ataupun aliran sungai yang tidak mampu menahan debit air hujan. … Hasil pengujian dengan menggunakan metode blackbox menunjukan bahwa penginputan pada proses login, menambah data, mengubah data, dan menghapus data akan ditanggapi sistem dengan memberikan notifikasi sesuai perintah masing-masing halaman”	“aplikasi kawasan rawan banjir rekomendasi evakuasi data dem dasar jarak jurus informatika banjir salah bencana alam faktor alam rupa curah hujan kelereng tanah rendah muka air laut tanggul alir sungai tahan debit air hujan hujan masyarakat sulit banjir henti banjir daerah kena banjir banjir masyarakat ungsi dasar teliti bangun buah sistem judul aplikasi kawasan rawan banjir rekomendasi evakuasi data dem dasar jarak aplikasi tuju informasi … simulasi hasil uji metode blackbox penginputan proses login data ubah data hapus data tanggap sistem notifikasi sesuai perintah halaman"

3) Vektorisasi TF

Pada tahap ini, teks yang telah melalui proses *preprocessing* akan diubah ke dalam bentuk representasi numerik menggunakan metode pembobotan TF. Proses vektorisasi ini dilakukan menggunakan fungsi vectorize.fit_transform, yang mengkonversi token-token dalam dokumen menjadi vektor numerik berdasarkan frekuensi kemunculannya serta signifikansi kata tersebut terhadap koleksi dokumen. Namun, dalam konteks penelitian ini, proses identifikasi dilakukan terhadap satu dokumen yang akan dicocokkan dengan daftar kata kunci. Oleh karena itu, nilai IDF untuk setiap *term* dianggap konstan, yaitu dengan asumsi bahwa jumlah dokumen N=1 dan setiap *term* memiliki df(t)=1. Dengan asumsi tersebut, komponen IDF tidak mempengaruhi hasil pembobotan, sehingga bobot TF-IDF secara efektif hanya merepresentasikan nilai TF dari masing-masing token dalam dokumen.

Pada Tabel 9. dibawah ditunjukkan pasangan kata dan nilai bobot dari vektorisasi dengan TF yang diurutkan dari bobot yang tertinggi. Kata dan bobot inilah yang akan menjadi tolak ukur dalam identifikasi dokumen.

TABEL IX
CONTOH HASIL VEKTORISASI TF

Term	Bobot
banjir	0.560798
rawan	0.213637
informasi	0.200285
rawan banjir	0.200285
sistem	0.146876
…	…
analisis	0.013352
manfaat	0.013352
antisipasi	0.013352

4) Identifikasi Kelompok Keahlian dengan Keyword Matching

Tahap ini merupakan inti dari sistem yang bertujuan untuk mengidentifikasi kelompok keahlian suatu dokumen tugas akhir berdasarkan kesesuaian kontennya dengan kumpulan kata kunci yang telah ditentukan. Proses identifikasi dilakukan dengan *keyword matching*, yang menggabungkan representasi vektor TF dari dokumen dengan daftar kata kunci tetap (*fixed keywords*) dan kata kunci dinamis (*dynamic keywords*) yang telah melalui proses *preprocessing*. Pada tahap ini, dokumen yang telah direpresentasikan dalam bentuk vektor TF akan dibandingkan dengan daftar kata kunci dari masing-masing kelompok keahlian. Setiap kata kunci yang cocok dengan token dalam dokumen akan diberikan bobot tertentu berdasarkan jenis kecocokan, yaitu *exact match, partial match* dan *match* pada judul.

TABEL X
CONTOH HASIL PERHITUNGAN KEYWORD MATCHING PADA KELOMPOK KEAHLIAN INFORMATION SYSTEM & DATA SPATIAL

Kata Kunci	Jenis Pencocokkan	Bobot TF	Match Judul	Skor Akhir
peta	Exact	0.0668	Tidak	0.2003
sistem informasi	Exact	0.0401	Tidak	0.1202
website	Exact	0.0134	Tidak	0.0401
informasi	Exact	0.2003	Tidak	0.6009
Total Skor				0.9614

Tabel 10. memperlihatkan rincian kontribusi kata kunci terhadap skor akhir setiap kelompok keahlian. Sebagai contoh, kata "informasi" memiliki nilai TF tertinggi yaitu 0.2003, dan setelah dikalikan dengan bobot *exact match*, menghasilkan skor akhir sebesar 0.6009. Kata lain seperti "peta" dan "sistem informasi" juga memberikan kontribusi positif masing-masing sebesar 0.2003 dan 0.1202.

TABEL XI
CONTOH HASIL PERHITUNGAN KEYWORD MATCHING PADA KELOMPOK KEAHLIAN SOFTWARE ENGINEERING & MOBILE COMPUTING

Kata Kunci	Jenis Pencocokkan	Bobot TF	Match Judul	Skor Akhir
aplikasi	Exact	0.0934	Ya	0.3645
Total Skor				0.3645

Sementara itu, pada Tabel 11. untuk kelompok *Software Engineering & Mobile Computing*, kata kunci "aplikasi" ditemukan pada judul dokumen, sehingga mendapatkan tambahan boost skor untuk *match* judul, menghasilkan nilai akhir sebesar 0.3645. Mekanisme boosting ini dirancang untuk memberikan penekanan lebih pada kecocokan kata kunci yang terdapat dalam judul, karena dianggap lebih merepresentasikan topik utama dokumen.

TABEL XII
CONTOH HASIL PERHITUNGAN KEYWORD MATCHING PADA KELOMPOK KEAHLIAN COMPUTATION & ARTIFICIAL INTELLIGENCE

Kata Kunci	Jenis Pencocokkan	Bobot TF	Match Judul	Skor Akhir
analisis	Exact	0.0133	Tidak	0.0401
rekomendasi	Exact	0.0801	Ya	0.3124
Total Skor				0.3525

Pada Tabel 12. untuk kelompok *Computation & Artificial Intelligence*, terdapat dua kata kunci yang teridentifikasi, yaitu "analisis" dan "rekomendasi". Kata kunci "analisis" memiliki bobot TF sebesar 0.0133 dan tidak ditemukan pada bagian judul, sehingga tidak mendapatkan tambahan skor, menghasilkan skor akhir sebesar 0.0401. Sementara itu, kata kunci "rekomendasi" ditemukan pada bagian judul, sehingga memperoleh *boosting* skor.

TABEL XIII
CONTOH HASIL PERHITUNGAN KEYWORD MATCHING PADA KELOMPOK KEAHLIAN NETWORKING & SECURITY

Kata Kunci	Jenis Pencocokkan	Bobot TF	Match Judul	Skor Akhir
aman	Exact	0.0267	Ya	0.0801
Total Skor				0.0801

Dan yang terakhir pada Tabel 13. dapat dilihat untuk kelompok *Networking & Security* hanya kata kunci “aman” yang ditemukan sehingga mendapat skor akhir 0.0801.

Kelompok keahlian *Information System & Data Spatial* memperoleh skor tertinggi sebesar 0.9614, diikuti oleh *Software Engineering & Mobile Computing* 0.3645, *Computation & Artificial Intelligence* 0.3525, dan *Networking & Security* 0.0801. Nilai tertinggi ini menjadi dasar bagi sistem untuk mengidentifikasi dokumen ke dalam kelompok *Information System & Data Spatial.*

C. Hasil Implementasi Pencarian Dokumen Serupa dengan Cosine Similarity

Setelah sebuah dokumen berhasil diidentifikasi ke dalam salah satu kelompok keahlian, sistem kemudian secara otomatis mencari dokumen lain dalam kelompok yang sama berdasarkan kemiripan isi menggunakan metode *cosine similarity*. Informasi dari dokumen serupa diharapkan dapat digunakan untuk validasi hasil identifikasi, memberikan pengguna konteks tambahan dari dokumen yang serupa.

TABEL XIV
CONTOH HASIL PENCARIAN DOKUMEN SERUPA DENGAN COSINE SIMILARITY

No	Judul Dokumen Serupa	Nilai
1	TATA KELOLA TEKNOLOGI INFORMASI MENGGUNAKAN FRAMEWORK COBIT 2019 (STUDI KASUS: UPT TIK UNIVERSITAS TANJUNGPURA PONTIANAK)	0.03
2	SISTEM INFORMASI MANAJEMEN PERMINTAAN DAN STOK DARAH (STUDI KASUS PMI KOTA PONTIANAK)	0.03
3	APLIKASI MARKETING PERUMAHAN MULTI USER DENGAN MENGGUNAKAN METODE LOCATION BASED SERVICE (STUDI KASUS: DPD REI PROVINSI KALIMANTAN BARAT)	0.02

Tabel 14. merupakan contoh hasil penerapan dari fitur pencarian dokumen serupa berbasis *cosine similarity*. Contoh hasil pada tabel ini menunjukkan daftar dokumen yang memiliki tingkat kemiripan konten tertinggi terhadap dokumen input, yang sebelumnya telah teridentifikasi sebagai bagian dari kelompok keahlian *Information System*.

D. Hasil Implementasi Kata Kunci Dinamis

Agar sistem tetap adaptif terhadap topik-topik baru, digunakan mekanisme memperbarui kata kunci dinamis dari dokumen-dokumen yang telah diidentifikasikan sebelumnya. Setelah pengguna mengklik *update* kata kunci dinamis, seluruh teks dokumen dalam satu kelompok digabungkan dan diolah menggunakan TF-IDF *vectorizer* untuk menghitung bobot *term* terhadap seluruh dokumen pada kelompok tersebut. Nilai bobot dihitung untuk setiap *term*, dan *term-term* dengan bobot tertinggi dipilih sebagai kata kunci dinamis. Sistem akan memilih sebanyak lima *term* teratas untuk masing-masing kelompok berdasarkan bobot TF-IDF tertinggi, yang kemudian disimpan dan digunakan dalam proses identifikasi dokumen selanjutnya sebagai kata kunci dinamis. Namun, sebelum *term-term* tersebut disimpan, sistem akan terlebih dahulu melakukan proses pengecekan terhadap daftar kata kunci tetap yang sudah ada untuk kelompok keahlian yang bersangkutan. Jika *term* yang dihasilkan dari perhitungan TF-IDF ternyata sudah ada dalam daftar kata kunci tetap, maka *term* tersebut akan dilewati dan sistem akan mengambil *term* dengan peringkat *term* TF-IDF berikutnya yang belum tercantum dalam kata kunci tetap.

Jika suatu kelompok keahlian belum memiliki cukup dokumen teridentifikasi, maka sistem tidak akan memperbarui kata kunci dinamis dan tetap mengosongkan daftar kata kunci dinamis.

TABEL XV
CONTOH HASIL KATA KUNCI DINAMIS

Kelompok Keahlian	Kata Kunci Dinamis Yang Dihasilkan
Computation & Artificial Intelligence	metode, sistem, bahasa, data, rupa
Networking & Security	-
Information System & Data Spatial	data, pontianak, uji, rupa, metode
Software Engineering & Mobile Computing	teliti, rancang, metode, kembang, user

Tabel 15. diatas menunjukkan hasil pembaruan kata kunci dinamis untuk masing-masing kelompok keahlian berdasarkan perhitungan bobot TF-IDF tertinggi dari dokumen-dokumen yang telah teridentifikasi sebelumnya.

E. Evaluasi Sistem

Untuk menilai performa sistem identifikasi dokumen dilakukan evaluasi menggunakan *confusion matrix*. Evaluasi ini dilakukan dengan membandingkan hasil identifikasi sistem

terhadap label sebenarnya (*ground truth*) dari dokumen yang telah diketahui kelompok keahliannya. Evaluasi ini bertujuan untuk mengukur kemampuan sistem dalam mengidentifikasi dokumen tugas akhir ke dalam empat kelompok keahlian yang telah didefinisikan.

TABEL XVI
METRIK EVALUASI

Label	Precision
Computation & Artificial Intelligence	0.933
Information System & Data Spatial	0.897
Networking & Security	1.000
Software Eng. & Mobile Computing	0.800
Accuracy	0.895

Berdasarkan Tabel 16 dapat dilihat akurasi keseluruhan sistem mencapai 89,5%, yang menunjukkan bahwa lebih dari 9 dari 10 dokumen berhasil dikelompokkan dengan benar. Dari hasil evaluasi tersebut dapat disimpulkan bahwa model dari sistem identifikasi dokumen tugas akhir yang dikembangkan mampu bekerja dengan akurat dan andal.

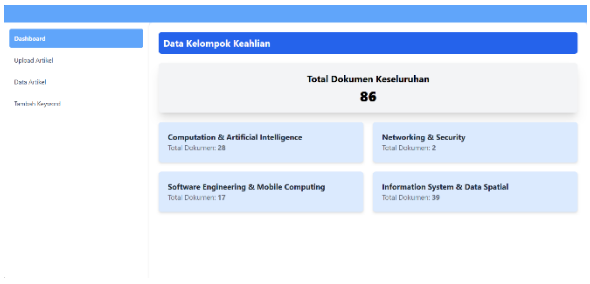
Meskipun masih terdapat beberapa kesalahan identifikasi yang setelah dianalisis disebabkan oleh dokumen yang diidentifikasi termasuk dalam lebih dari satu kategori, yang disebut sebagai identifikasi multilabel dan proses *document parsing* yang tidak sempurna, hasilnya tetap menunjukkan bahwa pendekatan TF dan *keyword matching* yang digunakan dalam penelitian ini efektif dalam mengenali kelompok keahlian dokumen berdasarkan konten utamanya.

F. Hasil Prototyping

Setelah seluruh proses perancangan dan pengembangan sistem dilakukan, tahap selanjutnya adalah implementasi prototipe sistem identifikasi dokumen tugas akhir ke dalam bentuk aplikasi berbasis web. Prototipe ini dibangun berdasarkan metode dan rancangan yang telah dijelaskan sebelumnya, dengan tujuan mengintegrasikan seluruh komponen utama.

1) Tampilan Antarmuka Sistem

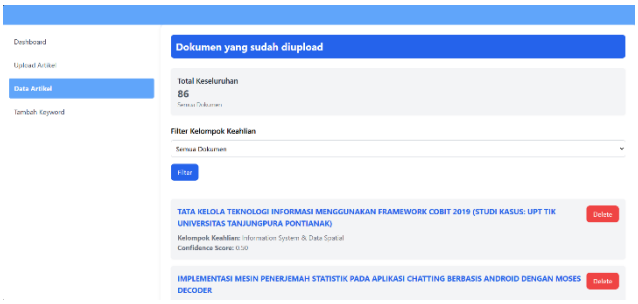
Tampilan antarmuka sistem terdiri dari empat halaman yang berfungsi menerima input dan menampilkan informasi. Adapun tampilan antarmuka sistem disajikan pada Gambar 9. hingga Gambar.



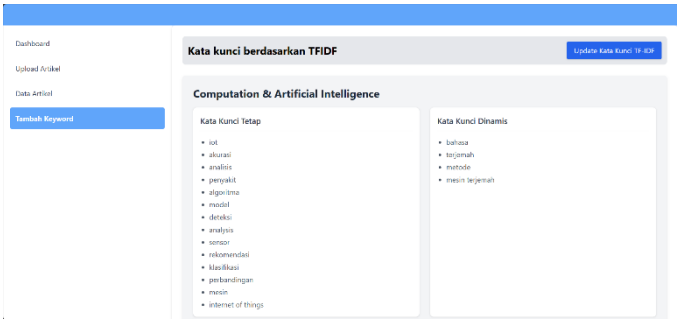
Gambar 9. Halaman Dashboard

Gambar 10. Halaman upload dokumen

Pada yang ditampilkan Gambar 10. pengguna dapat mengupload dokumen yang akan diidentifikasikan kelompok keahliannya. Setelah sistem berhasil melakukan semua proses akan ditampilkan infomasi berupa hasil identifikasi yang mencakup judul, *confidence score*, dan kelompok keahlian teridentifikasi. Sistem juga akan menampilkan hasil *parsing* dari dokumen serta nilai *confidence score* dan kata kunci yang cocok dari tiap tiap kelompok keahlian.



Gambar 11. Halaman data dokumen



Gambar 12. Halaman kata kunci

Gambar 12. adalah halaman kata kunci, dimana pengguna dapat melihat kata kunci tetap dan kata kunci dinamis setiap kelompok keahlian, user dapat meng*update* kata kunci secara berkala.

2) Blackbox testing

Tahapan pengujian sistem dilakukan untuk memastikan bahwa seluruh fungsi yang dikembangkan pada sistem identifikasi dokumen tugas akhir berjalan dengan baik sesuai dengan kebutuhan pengguna. Pengujian ini dilakukan menggunakan pendekatan *black box testing*, yaitu metode pengujian yang menilai kinerja sistem dari sisi fungsionalitas tanpa memperhatikan struktur internal kode.

Black box testing dilakukan dengan memberikan berbagai jenis input kepada sistem dan mengevaluasi apakah output yang dihasilkan telah sesuai dengan ekspektasi yang diharapkan. Tabel 17. merupakan hasil pengujian yang dilakukan terhadap sistem.

TABEL XVII
HASIL BLACKBOX TESTING

No	Pengujian	Input Uji	Ekspektasi Output	Status
1	Tampil Dashboard	Pengguna membuka halaman utama	Sistem menampilkan ringkasan statistik dan menu navigasi utama	Berhasil
2	Upload Dokumen	File .pdf atau .docx valid	Dokumen berhasil diunggah dan ditampilkan untuk identifikasi	Berhasil
3	Upload Dokumen (invalid)	File .jpg, atau .xlsx	Sistem menolak file dan menampilkan pesan kesalahan	Berhasil
4	Identifikasi Dokumen	Dokumen hasil unggah	Sistem memproses dan menampilkan judu, hasil identifikasi, confidence score dan daftar dokumen yang mirip	Berhasil
5	Tampil Seluruh Dokumen	Klik menu "Daftar Dokumen"	Sistem menampilkan seluruh dokumen yang sudah diidentifikasi	Berhasil
6	Filter Dokumen	Pilih filter berdasarkan kelompok keahlian	Sistem menampilkan dokumen sesuai dengan filter yang dipilih	Berhasil
7	Delete dokumen	Klik tombol delete pada suatu dokumen	Sistem menghapus dokumen	Berhasil
8	Tampil Kata Kunci	Klik menu "Daftar Kata Kunci"	Sistem menampilkan daftar kata kunci tetap dan dinamis	Berhasil
9	Update Kata Kunci Dinamis	Tersedia cukup dokumen untuk setiap kelompok	Sistem menghasilkan kata kunci dinamis berbasis TF-IDF	Berhasil
10	Update Kata Kunci Dinamis	Belum ada dokumen untuk kelompok tertentu	Sistem menampilkan notifikasi atau tidak melakukan update	Berhasil
11	Edit Kata Kunci Tetap	Kata kunci tetap sebanyak tepat 15 kata kunci	Sistem memperbarui kata kunci tetap sesuai input	Berhasil
12	Edit Kata Kunci Tetap	Kata kunci tetap tidak tepat sebanyak 15 kata kunci	Sistem menampilkan notifikasi kata kunci harus 15 kata kunci	Berhasil

IV. KESIMPULAN

Berdasarkan uraian yang telah disampaikan, dapat ditarik kesimpulan bahwa sistem yang dikembangkan mampu mengidentifikasi dokumen tugas akhir ke dalam empat kelompok keahlian dengan memanfaatkan pembobotan *Term Frequency* (TF) dan metode *keyword matching*. Pendekatan ini menggabungkan kata kunci tetap dan kata kunci dinamis yang dihasilkan secara otomatis, serta menerapkan skema pembobotan berdasarkan tingkat kecocokan, yaitu exact match, partial match, dan kemunculan kata kunci pada judul. Selain itu, sistem dilengkapi dengan fitur pencarian dokumen serupa berbasis *cosine similarity* yang bekerja pada representasi vektor TF-IDF, sehingga mampu mengukur kedekatan konten antar dokumen dan memastikan relevansi pencarian dalam kelompok keahlian yang sama. Evaluasi menggunakan *confusion matrix* menunjukkan akurasi sebesar 89,5%, yang menandakan sistem memiliki kinerja yang baik meskipun terdapat ketidakseimbangan jumlah data pada setiap

kelas. Temuan ini membuktikan bahwa kombinasi metode TF dan *keyword matching* efektif diterapkan untuk identifikasi dokumen berbasis teks di lingkungan akademik.

REFERENSI

[1] H. B. Modouw and P. I. Nugroho, "Faktor-Faktor Yang Mempengaruhi Penyelesaian Tugas Akhir Mahasiswa Program Studi Akuntansi Universitas Kristen Satya Wacana," J. Akunt. Profesi, vol. 12, no. 1, p. 59, 2021, doi: 10.23887/jap.v12i1.33283.

[2] E. R. N. Mustaqim, U. Pagalay, and ..., "Prediksi Tingkat Kepercayaan Masyarakat Terhadap Pilpres 2024 Menggunakan Tf-Idf Dan Bow Menggunakan Metode Svm," Mandalika ISSN 2721 …, pp. 515–530, 2024, [Online]. Available: https://www.ojs.cahayamandalika.com/index.php/jcm/article/view/3114%0Ahttps://www.ojs.cahayamandalika.com/index.php/jcm/article/download/3114/2521

[3] A. L. Tkachenko and L. A. Denisova, "Designing an information system for the electronic document management of a university : automatic classification of documents Designing an information system for the electronic document management of a university : automatic classification of documents," 2022, doi: 10.1088/1742-6596/2182/1/012035.

[4] M. Jurafsky, "Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models Third Edition draft Summary of Contents," Women's Hum. Rights, pp. vii–x, 2014.

[5] 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

[6] Nilawati, Husaini, and J. Salat, "Penggunaan Metode Cosine Similarity Dan Tf-Idf Untuk Klasifikasi Judul Seminar Proposal Pada Fakultas Teknik Universitas Jabal Ghafur," Sagita Acad. J., vol. 2, no. 1, pp. 72–79, 2024, doi: 10.61579/sagita.v2i1.60.

[7] M. Das, S. Kamalanathan, and P. Alphonse, "A Comparative Study on TF-IDF feature weighting method and its analysis using unstructured dataset," CEUR Workshop Proc., vol. 2870, pp. 98–107, 2021.

[8] Q. Zhang et al., "Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Data Extraction," 2024.

[9] A. Esra, A. Doa, A. Ayoub, F. Hussein, and S. Moh, "The Effect of Preprocessing Techniques , Applied to Numeric Features, on Classification Algorithms' Performance," Data, vol. 6, no. 11, 2021.

[10] A. Fatima, N. Nazir, and M. G. Khan, "Data Cleaning In Data Warehouse : A Survey of Data Pre-processing Techniques and Tools," no. March, pp. 50–61, 2017, doi: 10.5815/ijitcs.2017.03.06.

[11] J. Luis, G. John, L. Malagutti, and B. Dusell, "The Foundations of Tokenization : Statistical and Computational Concerns," pp. 1–15, 2024.

[12] Z. Abidin, A. Junaidi, and Wamiliana, "Text Stemming and Lemmatization of Regional Languages in Indonesia: A Systematic Literature Review," J. Inf. Syst. Eng. Bus. Intell., vol. 10, no. 2, pp. 217–231, 2024, doi: 10.20473/jisebi.10.2.217-231.

[13] A. Zubiaga, "TF-CR: Weighting Embeddings for Text Classification," pp. 1–11, 2020, [Online]. Available: http://arxiv.org/abs/2012.06606

[14] G. Mustafa, M. Usman, L. Yu, M. T. afzal, M. Sulaiman, and A. Shahid, "Multi-label classification of research articles using Word2Vec and identification of similarity threshold," Sci. Rep., vol. 11, no. 1, pp. 1–20, 2021, doi: 10.1038/s41598-021-01460-7.

[15] S. Up, S. Jalolov, and F. Odilova Altiboyevna, "About Django," pp. 538–541, 2024.

[16] L. Stępień and H. Dróżdż, Fundamentals of Python programming Examples and practice exercises. 2024.