KOMBINASI LATENT SEMANTIC INDEXING DAN SUPPORT VECTOR MACHINE PADA KLASIFIKASI DOKUMEN AKREDITASI (STUDI KASUS : PASCASARJANA UNIVERSITAS NEGERI MEDAN)

Angga Warjaya, Mansur As*, Inna Muthmainnah, Sri Mulyana,
Said Iskandar Al Idrus, Arnita, Insan Taufik
Ilmu Komputer, Universitas Negeri Medan
Jl. William Iskandar Psr. V, Kenangan Baru, Kecamatan Percut Sei Tuan,
Kabupaten Deli Serdang, Sumatera Utara, Indonesia 20221
asmansur@unimed.ac.id

ABSTRAK

Pengelolaan dokumen akreditasi yang efisien menjadi tantangan utama dalam pendidikan tinggi akibat volume dokumen yang besar dan format yang bervariasi. Penelitian ini bertujuan untuk mengembangkan metode klasifikasi otomatis menggunakan kombinasi *latent semantic indexing* dan *support vector machine* guna meningkatkan akurasi dan efisiensi pengelolaan dokumen akreditasi. Akurasi dalam penelitian ini mengacu pada ketepatan sistem dalam mengidentifikasi kategori dokumen sesuai kriteria akreditasi, sementara efisiensi mencerminkan percepatan dan penyederhanaan proses klasifikasi dibandingkan dengan metode manual. Dataset terdiri dari 230 dokumen yang dikategorikan berdasarkan kriteria Lembaga Akreditasi Mandiri Kependidikan, dengan 115 dokumen untuk Kriteria 6 (Pendidikan) dan 115 dokumen untuk Kriteria 7 (Penelitian), kemudian dibagi menjadi data latih dan uji dengan rasio 60:40. Proses klasifikasi dilakukan melalui beberapa tahap, termasuk pre-processing teks, ekstraksi fitur semantik, serta optimasi parameter model untuk memperoleh hasil terbaik. Pengujian menunjukkan bahwa metode yang diusulkan mampu mencapai tingkat akurasi sebesar 91%, dengan validasi silang sebesar 94,21%. Hasil ini menunjukkan bahwa pendekatan yang digunakan efektif dalam mengotomatisasi klasifikasi dokumen akreditasi, sehingga dapat mempercepat proses evaluasi serta meningkatkan efisiensi manajemen dokumen dalam institusi pendidikan tinggi.

Kata kunci : *akreditasi, klasifikasi dokumen, latent semantic indexing, support vector machine*.

1. PENDAHULUAN

Kemajuan pesat dalam teknologi informasi dan digitalisasi telah menyebabkan pertumbuhan data yang eksponensial, termasuk di sektor pendidikan [1]. Institusi pendidikan tinggi kini menghadapi peningkatan signifikan dalam jumlah data, yang mencakup catatan administrasi, dokumen akademik, serta laporan formal seperti dokumen akreditasi. Transformasi digital ini tidak hanya menuntut penyimpanan data, tetapi juga pengelolaan dan pemanfaatan yang efektif. Salah satu tantangan utama terletak pada pengelolaan dokumen penting seperti laporan akreditasi, yang berperan krusial dalam penjaminan mutu dan penilaian kinerja akademik. Dokumen-dokumen ini tidak hanya berfungsi sebagai bukti standar pendidikan, tetapi juga menjadi dasar dalam pengambilan keputusan strategis [2]. Seiring dengan meningkatnya volume dan kompleksitas dokumen, pendekatan manajemen tradisional menjadi semakin tidak efektif. Oleh karena itu, diperlukan solusi yang lebih otomatis dan efisien untuk mengoptimalkan pengelolaan dokumen.

Klasifikasi dokumen tetap menjadi tantangan utama dalam manajemen informasi, terutama bagi dokumen dengan struktur yang kompleks dan konten yang beragam, seperti laporan akreditasi. Proses pengkategorian dokumen secara manual tidak hanya memakan waktu, tetapi juga rentan terhadap kesalahan manusia. Selain itu, dokumen akreditasi sering kali tidak memiliki struktur standar, mengandung berbagai bahasa, dan menggunakan terminologi yang beragam. Variasi ini memperumit identifikasi pola dan hubungan yang dapat mendukung proses klasifikasi. Mengingat tantangan tersebut, otomatisasi proses klasifikasi menjadi semakin penting untuk meningkatkan efisiensi dan akurasi. Kecerdasan buatan, khususnya pembelajaran mesin, telah muncul sebagai solusi yang efektif untuk mengatasi permasalahan ini [3]. Melalui pembelajaran mesin, sistem dapat menganalisis data dan secara mandiri mengategorikan dokumen dengan mengenali pola-pola mendasar [4]. Implementasi teknologi ini secara signifikan mengurangi waktu yang dibutuhkan dalam klasifikasi, memungkinkan institusi pendidikan untuk lebih fokus pada analisis berbasis data serta pengambilan keputusan strategis dengan presisi dan efisiensi yang lebih baik.

Latent Semantic Indexing (LSI) adalah teknik yang bekerja dengan mengidentifikasi hubungan semantik atau makna tersembunyi antara kata-kata dalam suatu dokumen. Dengan menggunakan *Singular Value Decomposition* (SVD), LSI dapat mengurangi dimensi sekumpulan dokumen teks, sehingga memungkinkan identifikasi hubungan yang lebih dalam antar kata yang sering kali tidak terlihat melalui pendekatan yang lebih sederhana [5]. Dalam konteks klasifikasi dokumen akreditasi, LSI dapat membantu mengekstrak informasi penting dari volume data yang besar dan beragam dengan mengidentifikasi pola serta hubungan semantik antar dokumen [6]. Dengan demikian, LSI memainkan peran krusial dalam meningkatkan efektivitas klasifikasi dokumen dengan

menghasilkan representasi teks yang lebih kaya dan mendalam [7].

Support Vector Machine (SVM) adalah salah satu teknik pembelajaran mesin yang paling populer dalam tugas klasifikasi dan regresi, terutama untuk data yang bersifat non-linear [8]. SVM bekerja dengan mencari *hyperplane* optimal yang memisahkan kelas data dengan margin terbesar, sehingga memungkinkan klasifikasi yang sangat akurat bahkan ketika menangani data yang beragam [9]. Dalam konteks klasifikasi dokumen, seperti dokumen akreditasi, SVM telah terbukti sangat efektif karena dapat membedakan dokumen yang sangat mirip tetapi memiliki konteks dan isi semantik yang berbeda. Selain itu, SVM telah banyak diterapkan dalam berbagai aplikasi klasifikasi teks, termasuk deteksi spam, analisis sentimen, dan pengenalan pola dalam dokumen formal [10].

Meskipun LSI dan SVM masing-masing memiliki keunggulan dalam klasifikasi dokumen, kombinasi kedua metode ini dapat memberikan pendekatan yang lebih kuat dan efektif. LSI berperan dalam mereduksi dimensi data teks dengan mengungkap representasi semantik yang lebih dalam, membantu mengatasi masalah seperti noise dalam data, ambiguitas, dan sinonimi yang sering menghambat klasifikasi teks [11]. Setelah LSI digunakan untuk mengekstrak fitur semantik yang lebih ringkas dan relevan dari teks, SVM dapat memproses fitur-fitur ini untuk mencapai klasifikasi yang lebih akurat dan presisi tinggi. Kombinasi ini sangat efektif karena LSI menyederhanakan data tanpa kehilangan makna semantik yang kritis, sehingga membantu SVM dalam menangani data dengan dimensi yang lebih rendah tetapi tetap mempertahankan informasi yang berharga.

Dalam konteks klasifikasi dokumen akreditasi, dimana teks sering kali panjang, tidak terstruktur, dan kaya akan istilah teknis, kombinasi LSI dan SVM menjadi sangat relevan. LSI mengungkap hubungan konseptual tersembunyi dalam dokumen, sementara SVM berfungsi sebagai pengklasifikasi optimal dalam membedakan kategori dokumen berdasarkan pola yang teridentifikasi. Pendekatan ini telah berhasil diterapkan dalam berbagai aplikasi, seperti klasifikasi makalah ilmiah, dokumen hukum, dan email spam, menunjukkan bahwa kombinasi LSI dan SVM dapat menghasilkan hasil klasifikasi yang lebih baik dibandingkan metode tunggal [12]. Oleh karena itu, penelitian ini bertujuan untuk memanfaatkan keunggulan kedua teknik secara sinergis guna meningkatkan efisiensi dan akurasi klasifikasi dokumen akreditasi di Program Pascasarjana Universitas Negeri Medan, yang terdiri dari berbagai jenis dokumen dengan konten yang sangat beragam.

Dalam beberapa tahun terakhir, Program Pascasarjana Universitas Negeri Medan telah menerapkan sistem penyimpanan dokumen akreditasi berbasis web melalui platform https://efilling-ppsunimed.id/. Sistem ini memungkinkan penyimpanan dan pengelolaan dokumen yang penting dalam proses akreditasi. Namun, meskipun sistem ini telah mendukung digitalisasi penyimpanan dokumen, masih belum ada mekanisme otomatis untuk mengklasifikasikan dokumen berdasarkan kriteria yang ditetapkan oleh lembaga akreditasi. Pengelolaan dokumen secara manual dalam hal pengkategorian dan klasifikasi masih menjadi tantangan utama.

Dengan memahami peran krusial klasifikasi dokumen akreditasi dalam mendukung proses penjaminan mutu yang efektif, penelitian ini bertujuan untuk mengembangkan pendekatan klasifikasi dengan mengintegrasikan LSI dan SVM. Penerapan metode ini diharapkan dapat meningkatkan efisiensi dan akurasi pengelolaan dokumen akreditasi di Program Pascasarjana Universitas Negeri Medan.

2. TINJAUAN PUSTAKA

2.1. *Machine Learning*

Machine Learning (ML) merupakan bagian dari kecerdasan buatan yang memungkinkan komputer untuk menganalisis data dan mengambil keputusan tanpa harus diprogram secara eksplisit. Algoritma *machine learning* diklasifikasikan menjadi tiga jenis: terawasi (*supervised learning*), tanpa pengawasan (*unsupervised learning*), dan penguatan (*reinforcement learning*). Dalam klasifikasi dokumen, *supervised learning* sering digunakan, di mana sistem dilatih menggunakan *dataset* berlabel untuk memprediksi kategori dokumen yang tidak diketahui [13]

2.2. *Natural Language Processing* (NLP)

Natural Language Processing (NLP) merupakan cabang ilmu komputer yang berfokus pada interaksi antara komputer dan bahasa manusia. Dalam konteks penelitian ini, NLP digunakan untuk memahami dan mengekstraksi makna dari teks dalam dokumen akreditasi. Teknik-teknik NLP mencakup *tokenisasi*, *stemming*, penghapusan *stopwords*, dan ekstraksi fitur teks lainnya [14].

2.3. *Latent Semantic Indexing* (LSI)

Latent Semantic Indexing (LSI) merupakan teknik berbasis reduksi dimensi yang digunakan untuk menangani masalah *polysemi* dan *synonymy* dalam teks. LSI bekerja dengan mengubah teks menjadi ruang semantik laten, yang merupakan representasi konsep-konsep dalam dokumen. Teknik ini didasarkan pada *Singular Value Decomposition* (SVD), yang memetakan dokumen ke dalam ruang vektor yang lebih kecil, di mana hubungan semantik antar kata dapat diidentifikasi [5].

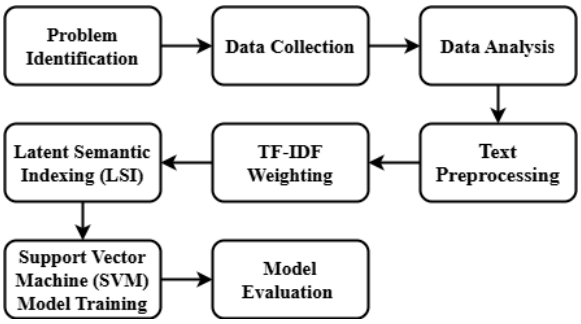
2.4. *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan algoritma pembelajaran mesin berjenis terawasi yang banyak digunakan untuk keperluan klasifikasi dan regresi. Cara kerja dari SVM adalah mencari *hyperplane* yang memisahkan data ke dalam dua kelas berbeda dengan margin terbesar. Dalam klasifikasi

teks, SVM sangat populer dari algoritma lainnya karena kemampuannya dalam menangani data berdimensi tinggi dan menghasilkan model yang kuat terhadap *overfitting*, terutama saat menghadapi data yang bersifat non-linear [10].

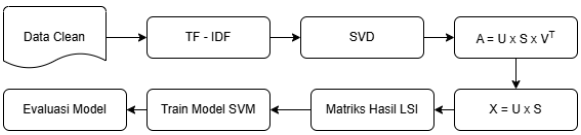
3. METODE PENELITIAN

Penelitian ini mengadopsi metodologi kuantitatif, yang berfokus pada pengukuran dan analisis data secara objektif menggunakan teknik numerik. Dalam penelitian ini, pendekatan kuantitatif diterapkan untuk mengevaluasi kinerja integrasi *Latent Semantic Indexing* (LSI) dengan *Support Vector Machine* (SVM) dalam klasifikasi dokumen akreditasi. Melalui metode ini, penelitian bertujuan untuk menilai akurasi sistem klasifikasi dan menentukan sejauh mana kombinasi LSI-SVM mampu mengelola dokumen akreditasi yang kompleks.



Gambar 1. Alur Penelitian

Alur penelitian merupakan kerangka kerja sistematis yang digunakan dalam pelaksanaan investigasi ilmiah. Bagian berikut menjelaskan tahapan-tahapan penelitian yang akan dilakukan, sebagaimana diilustrasikan pada Gambar 1.



Gambar 2. *Flowchart* Metode LSI dan SVM

3.1. Identifikasi Masalah

Proses akreditasi program studi di perguruan tinggi melibatkan sejumlah besar dokumen yang harus dikumpulkan, dikategorikan, dan dievaluasi berdasarkan kriteria tertentu. Hal ini sering menjadi tantangan yang signifikan, terutama dalam hal efisiensi waktu dan akurasi klasifikasi. Pengelolaan dokumen akreditasi secara manual membutuhkan upaya yang besar dan rentan terhadap kesalahan manusia. Tidak adanya sistem otomatis untuk mengklasifikasikan dokumen berdasarkan kriteria akreditasi memperlambat proses evaluasi dan penilaian. Oleh karena itu, diperlukan sistem klasifikasi otomatis yang dapat memproses dokumen akreditasi dengan cepat dan akurat, sehingga mempermudah pengelolaan dokumen dalam jumlah besar serta mempercepat proses akreditasi.

3.2. Pengumpulan Data

Data yang digunakan dalam penelitian ini berasal dari file PDF dokumen akreditasi yang diunduh melalui platform https://efilling-ppsunimed.id/. Setelah diunduh, file-file tersebut disimpan dalam direktori sesuai dengan kriteria yang ditetapkan oleh LAMDIK. Kriteria LAMDIK yang digunakan dalam penelitian ini meliputi Kriteria 6 - Pendidikan, yang mencakup kurikulum, metode pengajaran, dan penilaian pembelajaran, serta Kriteria 7 - Penelitian, yang mencakup kegiatan penelitian yang dilakukan oleh dosen dan mahasiswa, termasuk publikasi serta kontribusi ilmiah.

3.3. Analisis Data

Tahap analisis data bertujuan untuk memahami pola dan karakteristik data akreditasi yang digunakan dalam penelitian ini. Proses ini mencakup pemeriksaan struktur dan distribusi data, seperti distribusi kategori, kata kunci yang sering muncul, serta pola umum yang relevan dengan klasifikasi. Analisis dilakukan untuk mengidentifikasi fitur atau variabel penting yang dapat memengaruhi hasil klasifikasi serta memastikan kualitas data untuk tahap selanjutnya. Dengan demikian, analisis data akan membantu menentukan strategi pra-pemrosesan data yang lebih efektif, termasuk pemilihan teknik *Latent Semantic Indexing* (LSI) dan penentuan parameter optimal untuk model *Support Vector Machine* (SVM).

3.4. Pra-pemrosesan Teks

Tahap ini bertujuan untuk membersihkan dokumen akreditasi agar siap digunakan dalam proses pembobotan [15]. Pra-pemrosesan teks dalam klasifikasi teks melibatkan beberapa langkah yang harus dilakukan secara sistematis.

a. Konversi Format: Dokumen akreditasi dalam format PDF dikonversi menjadi teks biasa agar dapat dianalisis lebih lanjut.
b. *Case Folding*: Semua huruf dalam dokumen teks diubah menjadi huruf kecil untuk menghilangkan perbedaan antara huruf besar dan huruf kecil, sehingga analisis menjadi lebih konsisten. Misalnya, kata "Akreditasi" dan "akreditasi" akan diperlakukan sebagai kata yang sama setelah proses ini.
c. Penghapusan Tanda Baca: Semua tanda baca, simbol khusus, dan karakter yang tidak relevan seperti titik, koma, tanda seru, dan sebagainya dihapus dari teks. Langkah ini bertujuan agar hanya elemen yang relevan, yaitu kata-kata, yang dianalisis pada tahap selanjutnya.
d. Penghapusan *Stopword*: Kata-kata umum yang tidak memiliki kontribusi signifikan dalam analisis, seperti "dan", "atau", serta "yang" dihapus dari teks untuk meningkatkan efektivitas pemrosesan.

e. *Stemming*: Proses ini mengubah setiap kata menjadi bentuk dasarnya sesuai dengan konteksnya. Misalnya, kata "makan," "memakan," dan "dimakan" akan direduksi menjadi "makan.".
f. Tokenisasi: Teks yang telah diproses dipecah menjadi unit kata (token). Sebagai contoh, kalimat "Universitas Negeri Medan adalah institusi terakreditasi" akan diubah menjadi daftar token: ["universitas", "negeri", "medan", "adalah", "institusi", "terakreditasi"].

3.5. Pembobotan TF – IDF

Setelah tahap pra-pemrosesan selesai, dilakukan pembobotan menggunakan pendekatan *Term Frequency-Inverse Document Frequency* (TF-IDF). Langkah ini bertujuan untuk menentukan tingkat kepentingan setiap kata berdasarkan relevansinya dalam dokumen.

a. Perhitungan TF (*Term Frequency*): Menghitung frekuensi kemunculan setiap kata dalam sebuah dokumen untuk mengetahui seberapa sering kata tersebut muncul.
b. Perhitungan IDF (*Inverse Document Frequency*): Mengukur keunikan suatu kata dengan mempertimbangkan keberadaannya di seluruh kumpulan dokumen. Semakin jarang kata muncul dalam banyak dokumen, semakin tinggi nilai IDF-nya.
c. Pembentukan Matriks TF-IDF: Mengalikan nilai TF dan IDF untuk setiap kata guna menghasilkan matriks TF-IDF yang merepresentasikan bobot setiap kata dalam dokumen.

3.6. *Latent Semantic Indexing*

Setelah melewati tahap pra-pemrosesan, data akan memasuki tahap ekstraksi fitur menggunakan algoritma *Latent Semantic Indexing* (LSI) sebelum dilakukan proses klasifikasi. Algoritma LSI digunakan untuk mengidentifikasi kesamaan semantik antar kata dalam dokumen akreditasi.

a. Dekomposisi Matriks: Matriks TF-IDF diuraikan menggunakan metode *Singular Value Decomposition* (SVD) menjadi tiga komponen utama:
   - Matriks U: Mewakili dokumen dalam ruang topik.
   - Matriks S: Berisi nilai singular yang menunjukkan bobot pada setiap dimensi topik.
   - Matriks V^T: Mewakili kata-kata dalam ruang topik.
b. Seleksi Fitur: Fitur yang digunakan dalam proses klasifikasi SVM diperoleh dengan mengalikan *Matriks U* dan *Matriks S*.

3.7. Pelatihan Model *Support Vector Machine*

Setelah tahap ekstraksi fitur dengan *Latent Semantic indexing*, data akan memasuki tahap klasifikasi menggunakan algoritma *Support Vector Machine* (SVM). SVM merupakan algoritma pembelajaran mesin yang digunakan dalam penelitian ini untuk klasifikasi dokumen akreditasi. Algoritma ini bekerja dengan mencari *hyperplane* yang memisahkan data ke dalam kelas yang berbeda. Dalam penelitian ini, SVM dilatih menggunakan dataset yang diperoleh dari LSI untuk mengklasifikasikan dokumen akreditasi berdasarkan kriteria yang telah ditetapkan. Kemampuan SVM dalam menangani data berdimensi tinggi, seperti teks, menjadikannya pilihan yang ideal untuk klasifikasi dokumen yang kompleks. Proses pelatihan model SVM dimulai dengan data berdimensi rendah yang telah dihasilkan melalui LSI. Proses pelatihan ini terdiri dari beberapa langkah utama sebagai berikut:

a. Data Masukan
   - Data Pelatihan: Representasi dokumen yang diperoleh dari LSI digunakan sebagai data masukan untuk melatih model SVM.
   - Data Uji: Sebagian data (40% dari dataset) dipisahkan untuk mengevaluasi kinerja model setelah pelatihan.
b. Penerapan Fungsi Kernel
   Fungsi kernel digunakan untuk menentukan pola klasifikasi dalam data. Dalam penelitian ini, empat jenis kernel diuji:
   - Kernel Linear: Cocok untuk data yang dapat dipisahkan secara linear.
   - Kernel Polinomial: Memodelkan hubungan non-linear dengan derajat polinomial tertentu.
   - Kernel *Radial Basis Function* (RBF): Efektif untuk menangani hubungan non-linear yang kompleks.
   - Kernel Sigmoid: Meniru fungsi aktivasi sigmoid, cocok untuk pola tertentu dalam data.
c. Pelatihan SVM Secara Bertahap
   Pada tahap ini, model SVM dilatih secara iteratif menggunakan data pelatihan. Algoritma optimasi SVM akan mencari *hyperplane* optimal yang memaksimalkan margin antar kelas. Proses ini mencakup:
   - Menghitung bobot dan bias *hyperplane*.
   - Menyesuaikan parameter berdasarkan kernel yang dipilih.

3.8. Evaluasi Model

Setelah model *Support Vector Machine* (SVM) dilatih, kinerjanya dievaluasi menggunakan data uji untuk mengukur seberapa baik model tersebut mengklasifikasikan dokumen akreditasi berdasarkan kriteria yang telah ditetapkan. Metode evaluasi ini mencakup beberapa metrik utama, yaitu *accuracy*, *precision*, *recall*, dan *F1-score*. Metrik-metrik ini dihitung berdasarkan *Confusion Matrix*, yang membandingkan hasil klasifikasi yang dihasilkan oleh model dengan klasifikasi sebenarnya. Evaluasi ini bertujuan untuk memahami efektivitas kombinasi *Latent Semantic Indexing* (LSI) dan SVM dalam klasifikasi dokumen serta membandingkannya dengan

metode klasifikasi lainnya. Hasil evaluasi akan memberikan wawasan mengenai keunggulan dan keterbatasan model yang diusulkan, sehingga dapat ditentukan apakah pendekatan ini memberikan solusi yang lebih efisien dan akurat dalam klasifikasi dokumen *akreditasi*.

Tabel 1. Confusion Matrix

Nilai Prediksi	**Nilai Aktual**	
	Positive	***Negative***
Positive	*True Positive*	*False Negative*
Negative	*False Positive*	*True Negative*

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$recision = \frac{TP}{TP + FP} \quad (2)$$

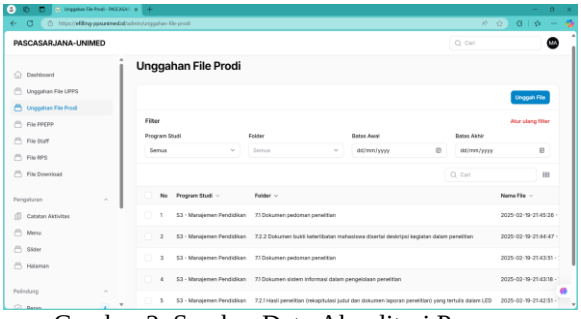
$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (4)$$

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Data dalam penelitian ini diperoleh dari platform https://efilling-ppsunimed.id/, yang dikelola oleh Program Pascasarjana Universitas Negeri Medan (PPs Unimed). Penelitian ini menggunakan 230 file dokumen akreditasi, yang terdiri dari 115 dokumen untuk Kriteria 6 (Pendidikan) dan 115 dokumen untuk Kriteria 7 (Penelitian). Seluruh dokumen akreditasi disimpan di Google Drive untuk memfasilitasi akses dan pemrosesan menggunakan Google Colab. Penyimpanan dokumen di Google Drive memungkinkan sistem untuk membaca dan memprosesnya secara efisien, terutama dalam proses ekstraksi teks untuk tugas klasifikasi.



Gambar 2. Sumber Data Akreditasi Program Pascasarjana Universitas Negeri Medan

4.2. Pra-pemrosesan Teks

Pada penelitian ini, tahap pra-pemrosesan teks dilakukan untuk membersihkan dan menyiapkan data sebelum proses klasifikasi. Langkah pertama adalah konversi dokumen akreditasi dari format PDF ke teks (.txt). Jika PDF berbasis teks, ekstraksi dilakukan menggunakan PyPDF2, sedangkan jika hasilnya kosong atau tidak terbaca, maka digunakan metode OCR dengan pytesseract setelah mengonversi halaman pertama PDF menjadi gambar menggunakan pdf2image. Setelah itu, dilakukan *case folding* dengan mengubah seluruh teks menjadi huruf kecil untuk memastikan konsistensi dalam analisis. Selanjutnya, tanda baca dihapus menggunakan pustaka string.punctuation agar tidak mengganggu proses pemrosesan teks. Kemudian, dilakukan penghapusan stopword menggunakan daftar *stopword* dari pustaka Sastrawi untuk menghilangkan kata-kata umum yang tidak memiliki makna signifikan. Setelah itu, teks melalui proses *stemming* dengan pustaka Sastrawi untuk mengubah kata berimbuhan menjadi bentuk dasarnya, sehingga mengurangi variasi kata yang tidak diperlukan. Langkah terakhir adalah tokenisasi menggunakan fungsi word_tokenize() dari pustaka NLTK untuk memisahkan teks menjadi kata-kata individu dalam bentuk list. Hasil dari masing-masing tahap preprocessing pada satu dokumen akreditasi ditunjukkan dalam tabel berikut:

Tabel 2. Pra-pemrosesan Teks

Tahap Pra-pemrosesan Teks	**Pembobotan TF-IDF**
Teks Awal	RENCANA PERKULIAHAN SEMESTER MATA KULIAH NEUROPSIKOLINGUISTIK UNIVERSITAS NEGERI MEDAN 2024.
Case Folding	rencana perkuliahan semester mata kuliah neuropsikolinguistik universitas negeri medan 2024.
Hapus Tanda Baca	rencana perkuliahan semester mata kuliah neuropsikolinguistik universitas negeri medan 2024
Hapus *Stopword*	rencana perkuliahan semester mata kuliah neuropsikolinguistik universitas negeri medan 2024
Stemming	rencana kuliah semester mata kuliah neuropsikolinguistik universitas negeri medan 2024
Tokenisasi	['rencana', 'kuliah', 'semester', 'mata', 'kuliah', 'neuropsikolinguistik', 'universitas', 'negeri', 'medan', '2024']

4.3. Pembobotan TF-IDF

Setelah tahap *text preprocessing* selesai, langkah selanjutnya adalah menerapkan pembobotan *TF-IDF* (*Term Frequency - Inverse Document Frequency*). Teknik ini digunakan untuk menilai tingkat kepentingan suatu kata dalam sebuah dokumen dengan menggabungkan dua konsep utama. *Term Frequency* (TF) mengukur seberapa sering suatu kata muncul dalam sebuah dokumen tertentu, sedangkan *Inverse Document Frequency* (IDF) memberikan bobot lebih besar pada kata-kata yang jarang muncul di seluruh koleksi dokumen. Dengan menerapkan pembobotan TF-IDF, dokumen dikonversi menjadi representasi vektor, di mana setiap dimensi merepresentasikan satu kata unik dalam dataset. Nilai dalam vektor ini

mencerminkan skor TF-IDF dari setiap kata dalam dokumen, yang menunjukkan tingkat kepentingannya. Untuk menghasilkan representasi vektor TF-IDF, penelitian ini menggunakan kelas TfidfVectorizer dari modul sklearn.feature_extraction.text dalam pustaka *scikit-learn*. Hasil representasi vektor TF-IDF untuk sebuah dokumen ditunjukkan pada ilustrasi berikut.

Tabel 3. Pembobotan TF-IDF

Pra-pemrosesan Teks	Pembobotan TF-IDF
['rencana', 'kuliah', 'semester', 'mata', 'kuliah', 'neuropsikolinguistik', 'universitas', 'negeri', 'medan', '2024']	2024: 0.260495, kuliah: 0.700510, mata: 0.350255, medan: 0.141839, negeri: 0.141839, neuropsikolinguistik: 0.350255, rencana: 0.260495, semester: 0.260495, universitas: 0.141839

4.4. Penerapan Latent Semantic Indexing (LSI)

Setelah menerapkan pembobotan TF-IDF, langkah berikutnya adalah mengimplementasikan *Latent Semantic Indexing* (LSI), yaitu metode yang digunakan untuk mengungkap hubungan laten antara kata-kata dalam dokumen. LSI menggunakan *Singular Value Decomposition* (SVD) untuk mengurangi dimensi vektor TF-IDF, sehingga informasi penting tetap terjaga dalam bentuk yang lebih ringkas dan bermakna. Dengan LSI, dokumen direpresentasikan sebagai vektor konsep, di mana setiap dimensinya tidak lagi secara langsung merujuk pada kata tertentu, tetapi pada konsep tersembunyi yang diperoleh dari pola hubungan antar kata dalam kumpulan dokumen. Teknik ini membantu mengurangi *noise* dan meningkatkan akurasi dalam analisis teks, terutama dalam tugas klasifikasi dan pencarian informasi. Dalam penelitian ini, proses LSI dilakukan menggunakan *TruncatedSVD* dari pustaka *scikit-learn*, yang memungkinkan ekstraksi sejumlah komponen utama dari data berbobot TF-IDF. Setiap dokumen kemudian direpresentasikan sebagai vektor dengan nilai yang mencerminkan hubungan laten dalam ruang konseptual. Representasi LSI yang dihasilkan untuk satu dokumen ditampilkan di bawah ini, di mana setiap dokumen direpresentasikan sebagai vektor konsep tunggal dengan nilai yang dipisahkan oleh tanda titik koma (";").

Tabel 4. Latent Semantic Indexing

Pembobotan TF-IDF	*Latent Semantic Indexing*
2024: 0.260495, kuliah: 0.700510, mata: 0.350255, medan: 0.141839, negeri: 0.141839, neuropsikolinguistik: 0.350255, rencana: 0.260495, semester: 0.260495, universitas: 0.141839	0.13826610589669738; 0.47038102052517033; -0.08962586292295172; 0.5213851113842708; 0.22747476208033532; 0.6185465406306543; 0.20760485809348828; -0.03560983002992534; -0.03203622134044191; -0.0031942019009835766

4.5. Pelatihan Model *Support Vector Machine*

Bagian ini menjelaskan proses pelatihan model *Support Vector Machine* (SVM) dalam klasifikasi dokumen akreditasi. Tujuan utama dari eksperimen ini adalah menentukan kombinasi parameter optimal untuk menghasilkan model dengan performa terbaik [16].

4.6. Penerapan Fungsi *Kernel*

Dalam penelitian ini, model SVM diuji dengan empat jenis *kernel*:

a. *Linear* – Cocok untuk data yang dapat dipisahkan secara *linier*, dengan keputusan klasifikasi berbasis pemisahan garis lurus.
b. *Radial Basis Function* (RBF) – Mampu menangani hubungan non-linier dengan mentransformasikan data ke dalam ruang berdimensi lebih tinggi.
c. *Sigmoid* – Mirip dengan fungsi aktivasi dalam jaringan saraf tiruan, digunakan untuk menangani data dengan pola distribusi kompleks.
d. *Polynomial* (*Poly*) – Menggunakan derajat polinomial, memungkinkan pembentukan batas keputusan yang lebih kompleks dibandingkan kernel linear.

Selain itu, dua hiperparameter utama dalam SVM juga divariasikan:

a. C (*Regularization Parameter*): Mengontrol keseimbangan antara margin maksimum dan minimisasi kesalahan. Nilai C yang kecil menghasilkan margin lebih luas tetapi dapat menyebabkan *underfitting*, sedangkan nilai C yang besar berfokus pada pemisahan ketat yang dapat menyebabkan *overfitting*.
b. Gamma (*Kernel Coefficient*, khusus untuk RBF, *Sigmoid*, dan *Poly*): Mengontrol kompleksitas fungsi keputusan, di mana nilai *gamma* tinggi menghasilkan batas keputusan lebih kompleks, sementara nilai *gamma* rendah menghasilkan model yang lebih generalisasi.

Eksperimen ini dilakukan sebanyak 100 kali dengan mengombinasikan 4 jenis kernel dan 10 kombinasi nilai C dan *gamma*. Proses pemilihan parameter terbaik dilakukan menggunakan *GridSearchCV* untuk menemukan parameter terbaik berdasarkan akurasi validasi silang (*cross-validation*).

4.7. Hasil dan Pemilihan Model Terbaik

Setelah melakukan 100 percobaan, hasil menunjukkan bahwa *kernel Sigmoid* dengan C = 0.1 dan *Gamma* = 10 memberikan performa terbaik dengan akurasi validasi silang tertinggi sebesar 94,21%. Berikut adalah ringkasan hasil terbaik untuk masing-masing kernel:

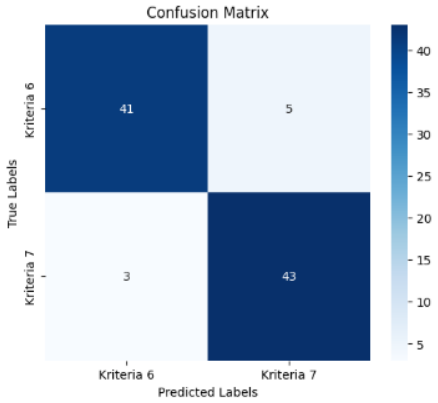
Tabel 5. Hasil Akurasi untuk Setiap Kernel

Kernel	**C**	***Gamma***	***Cross-Validation Score***
Sigmoid	0.1	10	0,942063*
Poly	0.1	10	0,898677
RBF	1	1	0,935185
Linear	1	0.1	0,927778

Hasil analisis menunjukkan bahwa kernel Sigmoid memberikan performa terbaik dibandingkan kernel lainnya karena kemampuannya dalam menangani pola hubungan kompleks yang muncul dari struktur semantik dokumen akreditasi. Penggunaan nilai gamma yang besar (10) memungkinkan model menangkap variasi yang lebih kaya dalam ruang fitur, sehingga lebih sensitif terhadap pola-pola laten yang tidak terdeteksi oleh pendekatan konvensional. Sementara itu, pemilihan nilai C = 0.1 menunjukkan bahwa regularisasi yang kuat, dengan margin pemisah yang lebih lebar, dapat mencegah *overfitting* pada *dataset* ini. Hal ini menjadi relevan karena representasi fitur hasil dari LSI telah menyederhanakan kompleksitas data, sehingga model tidak memerlukan pemisahan yang sangat ketat. Meskipun kernel RBF menunjukkan akurasi yang mendekati kernel Sigmoid, performanya masih berada di bawah karena lebih sensitif terhadap nilai gamma dan berisiko *overfitting* pada *gamma* tinggi. Di sisi lain, *kernel Linear* dan Polinomial memberikan hasil yang lebih rendah karena kurang fleksibel dalam memodelkan hubungan non-linier antar fitur dalam dokumen. Oleh karena itu, kombinasi *kernel Sigmoid* dengan C = 0.1 dan gamma = 10 memberikan keseimbangan optimal antara bias dan variansi, menghasilkan model klasifikasi yang stabil, presisi, dan efektif, menjadikannya pilihan terbaik dalam tugas klasifikasi dokumen akreditasi pada penelitian ini.

4.8. Evaluasi

Dalam penelitian ini, model klasifikasi dokumen dievaluasi menggunakan metrik *Confusion Matrix*, *Precision*, *Recall*, *F1-Score*, dan *Accuracy*. Hasil evaluasi menunjukkan bahwa model mampu mengklasifikasikan dokumen ke dalam dua kategori, yaitu Kriteria 6 dan Kriteria 7, dengan performa yang baik.



Gambar 3. *Confusion Matrix*

Untuk Kriteria 6, model memiliki nilai *Precision* sebesar 95%, yang berarti dari seluruh prediksi yang diklasifikasikan sebagai Kriteria 6, sekitar 95% di antaranya benar. Sementara itu, nilai *Recall* mencapai 87%, menunjukkan bahwa model berhasil mengidentifikasi 87% dari seluruh data aktual Kriteria 6. Nilai *F1-Score* sebesar 91% mencerminkan keseimbangan antara presisi dan sensitivitas model. Untuk Kriteria 7, model memiliki *Precision* sebesar 88% dan *Recall* sebesar 96%, dengan nilai *F1-Score* mencapai 92%. Hasil ini menunjukkan bahwa model secara konsisten dapat mendeteksi dokumen yang termasuk dalam kategori ini dengan tingkat keakuratan yang tinggi.

	precision	recall	f1-score	support
Kriteria 6	0.95	0.87	0.91	46
Kriteria 7	0.88	0.96	0.92	46
accuracy			0.91	92
macro avg	0.92	0.91	0.91	92
weighted avg	0.92	0.91	0.91	92

Gambar 4. *Classification Report*

Secara keseluruhan, model mencapai tingkat akurasi sebesar 91%, yang menunjukkan bahwa model dapat mengklasifikasikan dokumen dengan tingkat kesalahan yang rendah. Rata-rata nilai *Precision*, *Recall*, dan *F1-Score* juga menunjukkan hasil yang seimbang, dengan nilai di atas 91%. Temuan ini mengindikasikan bahwa pendekatan yang digunakan dalam klasifikasi dokumen akreditasi bersifat efektif dan andal.

5. KESIMPULAN DAN SARAN

Kombinasi *Natural Language Processing* (NLP) dengan teknik *Machine Learning*, khususnya *Latent Semantic Indexing* (LSI) dan *Support Vector Machine* (SVM), terbukti efektif dalam mengklasifikasikan dokumen akreditasi di Program Pascasarjana Universitas Negeri Medan (PPs Unimed) dengan tingkat akurasi yang tinggi. LSI mengubah teks dokumen menjadi representasi vektor berdimensi rendah, sehingga mampu menangkap hubungan semantik antar kata yang tidak dapat dideteksi oleh metode berbasis kata konvensional seperti TF-IDF. Sementara itu, SVM berperan sebagai model klasifikasi yang secara optimal memisahkan kelas dokumen. Proses optimasi menunjukkan bahwa pemilihan parameter dalam SVM memiliki dampak signifikan terhadap kinerja klasifikasi. Parameter utama yang berpengaruh adalah fungsi kernel, nilai *C* (regularisasi), dan parameter *gamma*. Hasil penelitian menemukan bahwa kombinasi parameter optimal yang menghasilkan akurasi tertinggi adalah kernel *Sigmoid* dengan *C* = 0.1 dan *gamma* = 10. Kombinasi ini memastikan keseimbangan antara *bias* dan *variance*, sehingga dapat mencegah *overfitting* maupun *underfitting*. Berdasarkan pengujian model, sistem klasifikasi dokumen akreditasi ini mampu mencapai akurasi sebesar 91%, precision rata-rata 91,5%, recall

91,5%, dan F1-score sebesar 91,5%. Hasil ini menunjukkan bahwa pendekatan kombinasi LSI dan SVM dapat digunakan secara efektif untuk pengklasifikasian dokumen akreditasi secara otomatis dan efisien.

DAFTAR PUSTAKA

[1] A. A. Fauzi *et al.*, *PEMANFAATAN TEKNOLOGI INFORMASI DI BERBAGAI SEKTOR PADA MASA SOCIETY 5.0*. Jambi: PT. Sonpedia Publishing Indonesia, 2023.

[2] M. Wali *et al.*, *PENERAPAN & IMPLEMENTASI BIG DATA DI BERBAGAI SEKTOR (Pembangunan Berkelanjutan Era Industri 4.0 dan Society 5.0)*. PT. Sonpedia Publishing Indonesia, 2023.

[3] A. Khaksar and H. Hassani, "Shifting from endangerment to rebirth in the Artificial Intelligence Age : An Ensemble Machine Learning Approach for Hawrami Text Classification," *arXiv Prepr. arXiv2409.16884.*, vol. 1, no. 1, pp. 1–19, 2024.

[4] J. D. Venkatesh, A. Jaiswal, and G. Nanda, "Comparing human text classification performance and explainability with large language and machine learning models using eye-tracking," *Sci. Rep.*, vol. 14, no. 1, pp. 1–12, 2024, doi: 10.1038/s41598-024-65080-7.

[5] D. Lan, A. Tian, Y. Wang, and Y. Li, "An overview of the principle , algorithm improvement and application based on the theory of latent semantic indexing," *Acad. J. Comput. Inf. Sci.*, vol. 4, no. 5, pp. 71–75, 2021, doi: 10.25236/AJCIS.2021.040510.

[6] E. H. Fernando and H. Toba, "Pemanfaatan Latent Semantic Indexing untuk Mengukur Potensi Kerjasama Jurnal Ilmiah Lintas Universitas," *J. Tek. Inform. dan Sist. Inf.*, vol. 6, no. 3, pp. 489–503, 2020.

[7] N. Palah and T. Hidayati, "Implementasi Information Retrieval System Pada Aplikasi Pencarian File Dokumen Menggunakan Metode Latent Semantic Indexing," vol. 3, no. 7, pp. 1704–1713, 2024.

[8] I. S. Al-Mejibli, J. K. Alwan, and D. H. Abd, "The effect of gamma value on support vector machine performance with different kernels," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 5, pp. 5497–5506, 2020, doi: 10.11591/IJECE.V10I5.PP5497-5506.

[9] O. A. Montesinos López, A. Montesinos López, and J. Crossa, *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer, 2022. doi: 10.1007/978-3-030-89010-0.

[10] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alexandria Eng. J.*, vol. 60, no. 3, pp. 3401–3409, 2021, doi: 10.1016/j.aej.2021.02.009.

[11] A. K. Singh, G. Wadhwa, M. Ahuja, K. Soni, and K. Sharma, "ScienceDirect Android Malware Detection using LSI-based Reduced Opcode Android Malware Detection using LSI-based Reduced Opcode Feature Vector Feature Vector," *Procedia Comput. Sci.*, vol. 173, no. 2020, pp. 291–298, 2020, doi: 10.1016/j.procs.2020.06.034.

[12] J. H. G. D. C. Cavalcanti and J. Menyhárt, "LSI with Support Vector Machine for Text Categorization – a practical example with Python," *Int. J. Eng. Manag. Sci.*, vol. 6, no. 3, pp. 18–29, 2021, doi: 10.21791/ijems.2021.3.2.

[13] R. Ofori-Boateng, M. Aceves-Martins, N. Wiratunga, and C. F. Moreno-Garcia, *Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review*, vol. 57, no. 8. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10844-w.

[14] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, 2023, doi: 10.1007/s11042-022-13428-4.

[15] G. A. Saputri and D. Alita, "Analisis Sentimen Twitter Terhadap Pemindahan Ibu Kota Negara Menggunakan Support Vector Machine," *J. Inform. J. Pengemb. IT*, vol. 9, no. 3, pp. 213–223, 2024, doi: 10.30591/jpit.v9i3.6612.

[16] M. Shamsi and S. Beheshti, "Separability and scatteredness (S&S) ratio-based efficient SVM regularization parameter, kernel, and kernel parameter selection," *Pattern Anal. Appl.*, vol. 28, no. 33, 2025, doi: https://doi.org/10.1007/s10044-025-01411-2.