ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

ANALISIS PENGARUH VARIASI NILAI P PADA METODE MINKOWSKI DISTANCE DALAM MENENTUKAN KEMIRIPAN ABSTRAK SKRIPSI

Harlen Gilbert Simanullang, Arina Prima Silalahi✉, Nadyarni Natalis Caesarin Duha
Fakultas Ilmu Komputer, Universitas Methodist Indonesia, Medan, Indonesia
Email: primaarinasilalahi@gmail.com

DOI: https://doi.org/10.46880/jmika.Vol9No2.pp255-263

ABSTRACT

The Computer Science Study Program of Universitas Methodist Indonesia is faced with the challenge of verifying the authenticity of student theses which is still done manually. This study applies the Minkowski Distance method to analyze the level of similarity of thesis abstracts using one hundred samples. The preprocessing stage is carried out through five systematic steps: cleansing to remove non-alphabetic characters, case folding for letter standardization, tokenizing for text splitting, filtering for stopword elimination, and stemming to obtain root words, resulting in word vectors that are analyzed. The Minkowski Distance method is implemented with three parameter variations P = 3, P = 5, and P = 7, where the selection of parameters is based on differences in sensitivity to vector dimensions, the higher the P value, the greater the emphasis on significant differences between dimensions. The test results show that the parameter P = 7 provides the most optimal similarity measurement with the smallest distance of 3.84 for documents with the highest similarity. These findings contribute to the development of a more effective similarity detection system to maintain academic integrity.

Keyword: Minkowski Distance, Similarity Analysis, Preprocessing, Thesis Abstract, Text Mining.

ABSTRAK

Program Studi Ilmu Komputer Universitas Methodist Indonesia dihadapkan pada tantangan verifikasi keaslian skripsi mahasiswa yang masih dilakukan secara manual. Penelitian ini menerapkan metode Minkowski Distance untuk menganalisis tingkat kesamaan abstrak skripsi menggunakan seratus sampel. Tahap preprocessing dilakukan melalui lima langkah sistematis: cleansing untuk menghilangkan karakter non-alfabetik, case folding untuk standardisasi huruf, tokenizing untuk pemecahan teks, filtering untuk eliminasi stopwords, dan stemming untuk mendapatkan kata dasar, menghasilkan vektor kata yang dianalisis. Metode Minkowski Distance diimplementasikan dengan tiga variasi parameter P = 3, P = 5, dan P = 7, dimana pemilihan parameter didasarkan pada perbedaan sensitivitas terhadap dimensi vector semakin tinggi nilai P, semakin besar penekanan pada perbedaan signifikan antar dimensi. Hasil pengujian menunjukkan bahwa parameter P = 7 memberikan pengukuran kemiripan paling optimal dengan jarak terkecil 3,84 untuk dokumen dengan kesamaan tertinggi. Temuan ini berkontribusi pada pengembangan sistem deteksi kesamaan yang lebih efektif untuk menjaga integritas akademik.

Kata Kunci: Minkowski Distance, Analisis Kemiripan, Preprocessing, Abstrak Skripsi, Text Mining.

PENDAHULUAN

Plagiarisme adalah tindakan mengambil karya orang lain, termasuk kekayaan intelektual, dan menggunakannya dalam karyanya sendiri tanpa memberikan pengakuan atau mencantumkan sumber asli sebagai referensi (Risparyanto, 2020). Dalam dunia akademis, plagiarisme sering terjadi karena adanya kemiripan pada teks atau abstrak skripsi tanpa mencantumkan sumber aslinya. Hal ini menimbulkan keraguan terhadap kejujuran dan keaslian karya ilmiah, terutama jika penulis tidak menyebutkan referensi yang digunakan.

Skripsi merupakan sebuah paparan tulisan hasil penelitian yang membahas suatu permasalahan atau fenomena dalam bidang ilmu tertentu dengan menggunakan kaidah-kaidah yang berlaku (Kurnia Aini, 2022). Skripsi harus mematuhi prinsip-prinsip objektivitas, berlandaskan pada data yang solid, dan kesimpulan harus diambil melalui prosedur yang jelas dan logis untuk menghindari plagiarisme.

Penelitian-penelitian sebelumnya telah menggunakan berbagai metode untuk mendeteksi kemiripan teks, seperti Cosine Similarity untuk penentuan kemiripan antar skripsi (Lindang et al., 2022) dan algoritma Boyer-Moore untuk menentukan

tingkat kemiripan pada pengajuan judul skripsi (Ahmad et al., 2021). Namun, kedua metode tersebut memiliki keterbatasan dalam mengukur kemiripan teks secara komprehensif. Cosine Similarity hanya mengukur sudut antar vektor dan mengabaikan besaran magnitudenya, sehingga kurang sensitif terhadap perbedaan frekuensi kata. Sementara algoritma Boyer-Moore berfokus pada pencocokan string secara eksak dan kurang efektif untuk mengidentifikasi kemiripan kontekstual. Kesenjangan penelitian ini menjadi dasar eksplorasi metode alternatif yang dapat mengatasi keterbatasan tersebut.

Skripsi terdiri dari elemen-elemen krusial yang mencakup pendahuluan, tinjauan pustaka, metodologi, hasil penelitian, dan kesimpulan, dengan abstrak yang menjadi komponen vital merangkum semua bagian tersebut. Abstrak ini sering kali menjadi bacaan pertama bagi reviewer atau pembaca lainnya, menyajikan secara ringkas tujuan, metodologi, hasil utama, dan kesimpulan dari penelitian. Dengan meningkatnya jumlah mahasiswa dan skripsi yang diajukan setiap tahun, menjadi semakin sulit untuk memastikan bahwa setiap abstrak skripsi unik. Pendekatan tradisional untuk memeriksa keunikan sering kali tidak mencukupi, membutuhkan metode yang lebih efisien dan efektif.

Text Mining adalah proses pengumpulan informasi secara intensif menggunakan alat dan metode khusus yang digunakan untuk menganalisis data dan dokumen. Text Mining adalah bagian dari penambangan data dan dapat menganalisis data semi terstruktur (Word, PDF, kutipan teks) dan data tidak terstruktur (Kambey et al., 2020). Tujuannya adalah untuk memahami dan mengambil informasi berguna dari sumber data dengan mengidentifikasi dan mengeksplorasi pola bahasa yang unik. Dalam kasus text mining, sumber data yang digunakan adalah collection atau unstructured collection dan memerlukan kategorisasi untuk menemukan informasi sejenis (Kambey et al., 2020). Text mining adalah penerapan konsep dan teknik data mining. Namun, karena penambangan teks melibatkan pemrosesan data teks tidak terstruktur, penambangan teks memiliki lebih banyak tahapan dibandingkan penambangan data. Berdasarkan hal ini, kita memerlukan langkah pertama untuk menyiapkan data teks untuk diproses: preprocessing.

Preprocessing teks adalah serangkaian langkah atau teknik yang digunakan untuk membersihkan, menyiapkan, dan mengubah teks mentah menjadi format yang lebih sesuai untuk analisis teks atau pemrosesan bahasa alami (NLP). Pada umumnya, preprocessing data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh system (Vendyansyah & Pranoto, 2021). Tujuan dari preprocessing teks adalah untuk meningkatkan kualitas data teks, menghilangkan noise, dan memastikan bahwa teks siap digunakan dalam berbagai aplikasi analisis teks, termasuk analisis sentimen, klasifikasi teks, ekstraksi informasi, dan lain-lain.

Abstrak skripsi umumnya berbentuk teks. Pengelompokkan teks umumnya melibatkan data teks yang tidak terstruktur, maka solusi untuk menemukan pola yang diinginkan untuk dijadikan kunci pengelompokkan dapat digunakan teknik Text Mining

(Kurniana, I. R., Muhima, R.R., Wardana, S., Hakimah, 2021). Dengan menerapkan metode Minkowski Distance dengan penyesuaian nilai p, diharapkan akan meningkatkan akurasi dalam mendeteksi dan mengelompokkan kemiripan antar abstrak. Teknik ini memungkinkan evaluasi cepat dan akurat dari kemiripan antar abstrak, yang memfasilitasi identifikasi potensi plagiarisme sebelum skripsi diajukan.

Metode jarak Minkowski bertindak sebagai metrik penting untuk ruang vektor, berfungsi sebagai norma dalam ruang, mencakup bentuk umum jarak Euclidean dan Manhattan (Catur et al., 2023). Signifikansi pemilihan nilai p dalam Minkowski Distance terletak pada kemampuannya untuk menyesuaikan sensitivitas pengukuran terhadap variasi dimensi teks. Nilai p yang lebih rendah, seperti p=1 (Manhattan) atau p=2 (Euclidean), cenderung memberikan bobot yang sama pada semua dimensi, sementara nilai p yang lebih tinggi (p>3) akan memberikan penekanan lebih pada dimensi dengan perbedaan yang besar. Dalam konteks analisis kemiripan abstrak skripsi, penggunaan nilai p yang lebih tinggi memungkinkan identifikasi lebih tepat terhadap dokumen dengan kesamaan struktur konseptual, bukan hanya kesamaan penggunaan kata individual. Kontribusi penelitian ini adalah mengeksplorasi efek variasi nilai p (3, 5, dan 7) untuk mencari parameter optimal yang dapat meningkatkan akurasi deteksi kemiripan teks dalam konteks akademik.

Fokus penelitian ini terletak pada nilai variasi p yang disesuaikan untuk meningkatkan atau mengurangi sensitivitas pengukuran terhadap perubahan dimensi teks. Dengan fokus pada variasi nilai p, penelitian ini mengevaluasi bagaimana perubahan nilai ini mempengaruhi kemiripan antar abstrak skripsi dan menemukan nilai yang paling efektif untuk mencapai

kemiripan maksimal yang dapat digunakan untuk pengembangan sistem deteksi plagiarisme yang lebih handal di lingkungan akademik.

METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan text mining untuk menganalisis kemiripan abstrak skripsi di Program Studi Ilmu Komputer Universitas Methodist Indonesia. Metodologi yang digunakan terdiri dari beberapa tahapan sistematis yang dimulai dari pengumpulan data, preprocessing teks, pembobotan kata, perhitungan jarak Minkowski dengan variasi nilai p, dan analisis hasil. Berikut adalah penjelasan rinci tentang setiap tahapan yang dilakukan dalam penelitian ini.

Tahapan Penelitian

Penelitian ini dilaksanakan dalam lima tahapan utama sebagai berikut:

1. Pengumpulan Data
   Tahap awal penelitian melibatkan pengumpulan 100 sampel abstrak skripsi dari Program Studi Ilmu Komputer Universitas Methodist Indonesia tahun 2019-2024. Data abstrak ini dikumpulkan dari repositori skripsi kampus dan disimpan dalam format teks untuk pemrosesan lebih lanjut.
2. Preprocessing Data
   Tahap preprocessing dilakukan untuk membersihkan dan menyiapkan data teks agar siap dianalisis. Preprocessing data dalam penelitian ini meliputi lima proses berurutan:
   a. Cleansing adalah Proses menghilangkan karakter-karakter yang tidak diperlukan seperti tanda baca, simbol, angka, dan elemen non-tekstual lainnya.
   b. Case Folding adalah Proses mengubah seluruh teks menjadi huruf kecil untuk menstandarisasi format teks.
   c. Tokenizing adalah Proses pemecahan teks menjadi unit-unit kata individual untuk analisis lebih lanjut.
   d. Filtering adalah Proses penyaringan kata-kata dengan menghilangkan stopwords (kata-kata umum seperti "dan", "yang", "di", dll.) yang tidak memiliki nilai signifikan dalam analisis.
   e. Stemming adalah Proses mengubah kata-kata menjadi bentuk dasarnya dengan menghilangkan imbuhan, akhiran, dan awalan.
3. Pembobotan Kata dengan Term Frequency (TF)
   Setelah praproses, penelitian dilanjutkan dengan pembobotan frasa menggunakan Term Frequency (TF). TF menyatakan jumlah kemunculan kata t dalam dokumen d.Pendekatan yang digunakan adalah dengan menghitung bobot suatu kata berdasarkan jumlah kemunculannya dalam dokumen (Widaningrum et al., 2022). Perhitungan bobot term dilakukan dengan formula:

$$q = (tf) * (idf)$$

   Dimana:
   q : Nilai bobot term
   tf : Nilai term frequency (frekuensi kemunculan kata dalam dokumen)
   idf : Nilai inverse document frequency (nilai yang mencerminkan seberapa penting kata tersebut dalam keseluruhan koleksi dokumen)
   Hasil pembobotan ini kemudian digunakan untuk membentuk vektor kata yang merepresentasikan setiap abstrak skripsi.
4. Penghitungan Kemiripan dengan Metode Minkowski Distance
   Tahap utama dalam penelitian ini adalah menghitung kemiripan antar abstrak skripsi menggunakan metode Minkowski Distance dengan variasi nilai p. Jarak Minkowski merupakan metrik jarak umum yang mencakup baik jarak Euclidean maupun Manhattan, dengan parameter p yang dapat disesuaikan untuk pengukuran jarak yang lebih adaptif dan presisi (Euclidean & Dalam, 2024).
   Rumus untuk menghitung Minkowski Distance antara dua titik X = (x_1, x_2, ..., x_n) dan Y = (y_1, y_2, ..., y_n) dalam ruang berdimensi n adalah:

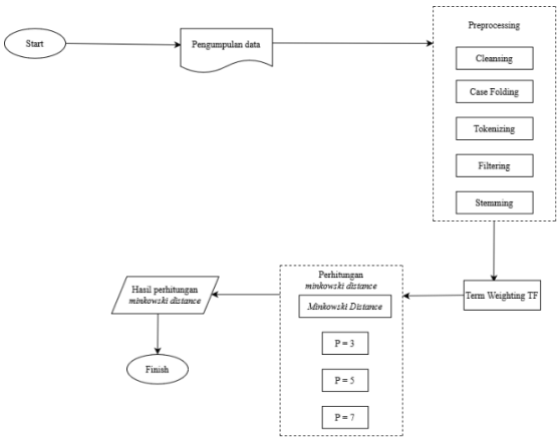
$$D(X,Y) = (\sum_{i=1}^{n} |xi - yi|^p)^{1/p}$$

   Dimana:
   d(X,Y) : Jarak Minkowski antara dua titik X dan Y
   n : Jumlah dimensi dari vektor X dan Y
   xi, yi : Komponen ke-i dari vektor X dan Y
   $|x_i - y_i|$: Nilai absolut dari perbedaan antar komponen ke-i
   p : Parameter yang menentukan tipe metrik
   Dalam penelitian ini, tiga nilai p yang berbeda diuji:
   a. p = 3: Memberikan sensitivitas moderat dalam pengukuran jarak
   b. p = 5: Meningkatkan penekanan pada dimensi dengan perbedaan besar
   c. p = 7: Memberikan penekanan maksimal pada dimensi dengan perbedaan signifikan
5. Analisis dan Evaluasi Hasil
   Tahap akhir penelitian adalah menganalisis dan mengevaluasi hasil perhitungan kemiripan dengan ketiga variasi nilai p. Analisis dilakukan dengan membandingkan dokumen uji (D100) dengan 99 dokumen lainnya (D1-D99) untuk setiap nilai p. Evaluasi difokuskan pada dokumen yang memiliki

jarak Minkowski terkecil, yang mengindikasikan tingkat kemiripan tertinggi dengan dokumen uji.

Desain Penelitian

Visualisasi desain penelitian dapat dilihat pada Gambar 1, yang menggambarkan alur proses dari pengumpulan data hingga analisis kemiripan menggunakan Minkowski Distance dengan variasi nilai P.



Gambar 1. Desain Penelitian

Keseluruhan metodologi penelitian ini dirancang untuk mengevaluasi efektivitas metode Minkowski Distance dengan berbagai nilai p dalam mendeteksi kemiripan abstrak skripsi, serta menemukan nilai p yang paling optimal untuk konteks penelitian ini.

IMPLEMENTASI DAN PENGUJIAN

Implementasi

Tahapan pengujian sistem menggunakan data abstrak skripsi dari Program Studi Ilmu Komputer Universitas Methodist Indonesia tahun 2019-2024, menggunakan 100 sampel abstrak skripsi.

Preprocessing Data

Dalam preprocessing data abstrak skripsi, teks telah dinormalisasi, dibersihkan dari karakter non-alfabetik dan simbol yang tidak diperlukan, serta telah dihilangkan stop words-nya. Dengan demikian, data diubah menjadi format yang lebih seragam dan bersih, mempersiapkannya untuk pembobotan dan analisis lebih lanjut.

Cleaning

Proses Cleaning melibatkan pembersihan teks dengan menghapus karakter-karakter yang tidak diperlukan seperti tanda baca dan simbol-simbol, sehingga tersisa hanya bagian teks yang memang dibutuhkan utuk di analisis. Contoh hasil cleaning dapat dilihat pada Tabel 1.

Tabel 1. Contoh Hasil Cleaning

Sebelum	Sesudah
Sistem Pendukung Keputusan adalah sebuah sistem yang dapat menyelesaikan permasalahan yang timbul dalam pemeringkatan dengan cepat dan dapat mengetahui nilai tertinggi hingga terendah dalam suatu percabangan. pada tulisan ini merupakan salah satu studi kasus yang dapat diselesaikan dengan menggunakan sistem pendukung keputusan, dimana yang menjadi persoalan yang di hadapi di SMK Swasta Teladan Medan adalah cara menyeleksi instruktur dan tenaga kerja yang unggul di dalam fakultas dan untuk melakukan berbagai macam harus menggunakan teknik panduan dan prosedur penilaian memerlukan waktu yang lama untuk mendapatkan konsekuensinya. oleh karena itu dibuatlah sebuah mesin bantu seleksi dengan menggunakan teknik Tsukamoto. untuk mengetahui keakuratan nilai yang diterima oleh alat tersebut, untuk bobot dan kriterianya.	Sistem Pendukung Keputusan adalah sebuah sistem yang dapat menyelesaikan permasalahan yang timbul dalam pemeringkatan dengan cepat dan dapat mengetahui nilai tertinggi hingga terendah dalam suatu percabangan pada tulisan ini merupakan salah satu studi kasus yang dapat diselesaikan dengan menggunakan sistem pendukung keputusan dimana yang menjadi persoalan yang di hadapi di SMK Swasta Teladan Medan adalah cara menyeleksi instruktur dan tenaga kerja yang unggul di dalam fakultas dan untuk melakukan berbagai macam harus menggunakan teknik panduan dan prosedur penilaian memerlukan waktu yang lama untuk mendapatkan konsekuensinya oleh karena itu dibuatlah sebuah mesin bantu seleksi dengan menggunakan teknik Tsukamoto untuk mengetahui keakuratan nilai yang diterima oleh alat tersebut untuk bobot dan kriterianya

Case Folding

Pada tahapan Case Folding, semua teks dikonversi menjadi huruf kecil untuk menghindari perbedaan perlakuan antara huruf besar dan kecil. Contoh hasil case folding dapat dilihat pada Tabel 2.

Tabel 2. Contoh Hasil Case Folding

Sebelum	Sesudah
Sistem Pendukung Keputusan adalah sebuah sistem yang dapat menyelesaikan permasalahan yang timbul dalam pemeringkatan dengan cepat dan dapat mengetahui nilai tertinggi hingga terendah dalam suatu percabangan. pada tulisan ini merupakan salah satu studi kasus yang dapat diselesaikan dengan menggunakan sistem pendukung keputusan dimana yang menjadi persoalan yang di hadapi di SMK Swasta Teladan Medan adalah cara menyeleksi instruktur dan tenaga kerja yang unggul di dalam fakultas dan untuk melakukan berbagai macam harus menggunakan teknik panduan dan prosedur penilaian memerlukan waktu yang lama untuk mendapatkan konsekuensinya oleh karena itu dibuatlah sebuah mesin bantu seleksi dengan menggunakan teknik Tsukamoto untuk mengetahui keakuratan nilai yang diterima oleh alat tersebut untuk bobot dan kriterianya	sistem pendukung keputusan adalah sebuah sistem yang dapat menyelesaikan permasalahan yang timbul dalam pemeringkatan dengan cepat dan dapat mengetahui nilai tertinggi hingga terendah dalam suatu percabangan pada tulisan ini merupakan salah satu studi kasus yang dapat diselesaikan dengan menggunakan sistem pendukung keputusan dimana yang menjadi persoalan yang di hadapi di smk swasta teladan medan adalah cara menyeleksi instruktur dan tenaga kerja yang unggul di dalam fakultas dan untuk melakukan berbagai macam harus menggunakan teknik panduan dan prosedur penilaian memerlukan waktu yang lama untuk mendapatkan konsekuensinya oleh karena itu dibuatlah sebuah mesin bantu seleksi dengan menggunakan teknik tsukamoto untuk mengetahui keakuratan nilai yang diterima oleh alat tersebut untuk bobot dan kriterianya

Tokenizing

Pada tahapan tokenizing, dilakukan pemecahan teks ke dalam unit-unit kecil berupa kata untuk memudahkan analisis di tahap berikutnya. Contoh hasil tokenizing dapat dilihat pada Tabel 3.

Tabel 3. Contoh Hasil Tokenizing

Sebelum	Sesudah
sistem pendukung keputusan adalah sebuah sistem yang dapat menyelesaikan permasalahan yang timbul dalam pemeringkatan dengan cepat dan dapat mengetahui nilai tertinggi hingga terendah dalam suatu percabangan pada tulisan ini merupakan salah satu studi kasus yang dapat diselesaikan dengan menggunakan sistem pendukung keputusan dimana yang menjadi persoalan yang di hadapi di smk swasta teladan medan adalah cara menyeleksi instruktur dan tenaga kerja yang unggul di dalam fakultas dan untuk melakukan berbagai macam harus menggunakan teknik panduan dan prosedur penilaian memerlukan waktu yang lama untuk mendapatkan konsekuensinya oleh karena itu dibuatlah sebuah mesin bantu seleksi dengan menggunakan teknik tsukamoto untuk mengetahui keakuratan nilai yang diterima oleh alat tersebut untuk bobot dan kriterianya	‘sistem’, ‘pendukung’, ‘keputusan’, ‘adalah’, ‘sebuah’, ‘sistem’, ‘yang’, ‘dapat’, ‘menyelesaikan’, ‘permasalahan’, ‘yang’, ‘timbul’, ‘dalam’, ‘pemeringkatan’, ‘dengan’, ‘cepat’, ‘dan’, ‘dapat’, ‘mengetahui’, ‘nilai’, ‘tertinggi’, ‘hingga’, ‘terendah’, ‘dalam’, ‘suatu’, ‘percabangan’, ‘pada’, ‘tulisan’, ‘ini’, ‘merupakan’, ‘salah’, ‘satu’, ‘studi’, ‘kasus’, ‘yang’, ‘dapat’, ‘diselesaikan’, ‘dengan’, ‘menggunakan’, ‘sistem’, ‘pendukung’, ‘keputusan’, ‘dimana’, ‘yang’, ‘menjadi’, ‘persoalan’, ‘yang’, ‘di’, ‘hadapi’, ‘di’, ‘smk’, ‘swasta’, ‘teladan’, ‘medan’, ‘adalah’, ‘cara’, ‘menyeleksi’, ‘instruktur’, ‘dan’, ‘tenaga’, ‘kerja’, ‘yang’, ‘unggul’, ‘di’, ‘dalam’, ‘fakultas’, ‘dan’, ‘untuk’, ‘melakukan’, ‘berbagai’, ‘macam’, ‘harus’, ‘menggunakan’, ‘teknik’, ‘panduan’, ‘dan’, ‘prosedur’, ‘penilaian’, ‘memerlukan’, ‘waktu’, ‘yang’, ‘lama’, ‘untuk’, ‘mendapatkan’, ‘konsekuensinya’, ‘oleh’, ‘karena’, ‘itu’, ‘dibuatlah’, ‘sebuah’, ‘mesin’, ‘bantu’, ‘seleksi’, ‘dengan’, ‘menggunakan’, ‘teknik’, ‘tsukamoto’, ‘untuk’, ‘mengetahui’, ‘keakuratan’, ‘nilai’, ‘yang’, ‘diterima’, ‘oleh’, ‘alat’, ‘tersebut’, ‘untuk’, ‘bobot’, ‘dan’, ‘kriterianya’

Filtering atau Stopword

Pada tahapan filtering, dilakukan penyaringan dengan menghapus kata-kata yang tidak memiliki makna dalam analisis, termasuk kata depan dan kata penghubung dari dalam teks. Contoh hasil filtering dapat dilihat pada Tabel 4.

Tabel 4. Contoh Hasil Filtering

Sebelum	Sesudah
'sistem', 'pendukung', 'keputusan', 'adalah', 'sebuah', 'sistem', 'yang', 'dapat', 'menyelesaikan', 'permasalahan', 'yang', 'timbul', 'dalam', 'pemeringkatan', 'dengan', 'cepat', 'dan', 'dapat', 'mengetahui', 'nilai', 'tertinggi', 'hingga', 'terendah', 'dalam', 'suatu', 'percabangan', 'pada', 'tulisan', 'ini', 'merupakan', 'salah', 'satu', 'studi', 'kasus', 'yang', 'dapat', 'diselesaikan', 'dengan', 'menggunakan', 'sistem', 'pendukung', 'keputusan', 'dimana', 'yang', 'menjadi', 'persoalan', 'yang', 'di', 'hadapi', 'di', 'smk', 'swasta', 'teladan', 'medan', 'adalah', 'cara', 'menyeleksi', 'instruktur', 'dan', 'tenaga', 'kerja', 'yang', 'unggul', 'di', 'dalam', 'fakultas', 'dan', 'untuk', 'melakukan', 'berbagai', 'macam', 'harus', 'menggunakan', 'teknik', 'panduan', 'dan', 'prosedur', 'penilaian', 'memerlukan', 'waktu', 'yang', 'lama', 'untuk', 'mendapatkan', 'konsekuensinya', 'oleh', 'karena', 'itu', 'dibuatlah', 'sebuah', 'mesin', 'bantu', 'seleksi', 'dengan', 'menggunakan', 'teknik', 'tsukamoto', 'untuk', 'mengetahui', 'keakuratan', 'nilai', 'yang', 'diterima', 'oleh', 'alat', 'tersebut', 'untuk', 'bobot', 'dan', 'kriterianya'	'sistem', 'pendukung', 'keputusan', 'menyelesaikan', 'permasalahan', 'timbul', 'pemeringkatan', 'cepat', 'mengetahui', 'nilai', 'tertinggi', 'terendah', 'percabangan', 'tulisan', 'merupakan', 'studi', 'kasus', 'diselesaikan', 'menggunakan', 'sistem', 'pendukung', 'keputusan', 'persoalan', 'smk', 'swasta', 'teladan', 'medan', 'cara', 'menyeleksi', 'instruktur', 'tenaga', 'kerja', 'unggul', 'fakultas', 'melakukan', 'berbagai', 'macam', 'teknik', 'panduan', 'prosedur', 'penilaian', 'memerlukan', 'waktu', 'mendapatkan', 'konsekuensinya', 'dibuatlah', 'mesin', 'bantu', 'seleksi', 'menggunakan', 'teknik', 'tsukamoto', 'mengetahui', 'keakuratan', 'nilai', 'diterima', 'alat', 'bobot', 'kriterianya'

Stemming

Di dalam stemming, setiap kata diubah menjadi kata sederhana untuk memperkecil jumlah frasa yang memiliki makna yang sama. Contoh hasil stemming dapat dilihat pada tabel 5.

Tabel 5. Contoh Hasil Stemming

Sebelum	Sesudah
'sistem', 'pendukung', 'keputusan', 'menyelesaikan', 'permasalahan', 'timbul', 'pemeringkatan', 'cepat', 'mengetahui', 'nilai', 'tertinggi', 'terendah', 'percabangan', 'tulisan', 'merupakan', 'studi', 'kasus', 'diselesaikan', 'menggunakan', 'sistem', 'pendukung', 'keputusan', 'persoalan', 'smk', 'swasta', 'teladan', 'medan', 'cara', 'menyeleksi', 'instruktur', 'tenaga', 'kerja', 'unggul', 'fakultas', 'melakukan', 'berbagai', 'macam', 'teknik', 'panduan', 'prosedur', 'penilaian', 'memerlukan', 'waktu', 'mendapatkan', 'konsekuensinya', 'dibuatlah', 'mesin', 'bantu', 'seleksi', 'menggunakan', 'teknik', 'tsukamoto', 'mengetahui', 'keakuratan', 'nilai', 'diterima', 'alat', 'bobot', 'kriterianya'	'sistem', 'dukung', 'putus', 'selesai', 'masalah', 'timbul', 'peringkat', 'cepat', 'tahu', 'nilai', 'tinggi', 'rendah', 'cabang', 'tulis', 'rupa', 'studi', 'kasus', 'selesai', 'guna', 'sistem', 'dukung', 'putus', 'soal', 'smk', 'swasta', 'teladan', 'medan', 'cara', 'seleksi', 'instruktur', 'tenaga', 'kerja', 'unggul', 'fakultas', 'laku', 'bagi', 'macam', 'teknik', 'panduan', 'prosedur', 'nilai', 'perlu', 'waktu', 'dapat', 'konsekuensi', 'buat', 'mesin', 'bantu', 'seleksi', 'guna', 'teknik', 'tsukamoto', 'tahu', 'akurat', 'nilai', 'terima', 'alat', 'bobot', 'kriteria'

Pembobotan Data

Setelah proses preprocessing diselesaikan, frekuensi kemunculan kata dalam setiap dokumen dihitung dengan menggunakan metode Term

Frequency (TF). Contoh penerapan pembobotan kata dapat dilihat pada Tabel 6.

No	*TERM*	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D100
1	a	1	0	0	0	0	0	0	0	0	0	0
2	abstrak	1	0	0	0	0	0	0	0	0	0	0
3	acu	0	0	1	0	0	0	0	0	0	0	0
4	admin	0	0	0	0	0	0	0	0	1	0	0
5	adopsi	0	0	0	1	0	0	0	0	0	0	0
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
402	video	0	0	0	0	1	0	0	0	0	0	0
403	virus	0	0	0	0	0	0	0	1	0	0	0
404	wabah	0	0	0	0	0	0	0	1	0	0	0
405	wortel	0	0	11	0	0	0	0	0	0	0	0
406	zona	0	0	0	0	0	0	0	3	0	0	0

Gambar *2*. Contoh Hasil Pembobotan TF

Penghitung Kemiripan Dokumen dengan Metode Minkowski Distance

Setelah preprocessing dan pembobotan data selesai, dilakukan analisis kemiripan antar abstrak skripsi menggunakan metode Minkowski Distance dengan variasi nilai P.

Perhitungan dengan P = 3

Hasil perhitungan kemiripan antara dokumen uji D100 dengan sepuluh dokumen pembanding menggunakan Minkowski Distance dengan P = 3 dapat dilihat pada Tabel 6.

Tabel 6. Hasil Perhitungan Minkowski Distance dengan P = 3

Data Uji	Pembanding	Distance
D100	D1	7.287362
D100	D2	11.36774
D100	D3	13.83131
D100	D4	7.34342
D100	D5	9.165656
D100	D6	13.70825
D100	D7	13.40908
D100	D8	9.113782
D100	D9	11.71553
D100	D10	11.17623

Dari tabel di atas, hasil perhitungan menggunakan Minkowski Distance dengan parameter P = 3 menunjukkan bahwa dokumen D1 memiliki jarak terkecil yaitu 7.287362, yang berarti memiliki tingkat kemiripan tertinggi dengan dokumen uji D100.

Perhitungan dengan P = 5

Hasil perhitungan kemiripan menggunakan Minkowski Distance dengan P = 5 dapat dilihat pada Tabel 7.

Tabel 7. Hasil Perhitungan Minkowski Distance dengan P = 5

Data Uji	Pembanding	Distance
D100	D1	4.99936
D100	D2	8.927007
D100	D3	11.57038
D100	D4	5.258298
D100	D5	7.33897
D100	D6	10.89707
D100	D7	10.27133
D100	D8	6.658982
D100	D9	8.496529
D100	D10	8.60793

Dari tabel di atas, hasil perhitungan menggunakan nilai P = 5 juga menunjukkan bahwa dokumen D1 memiliki jarak terkecil yaitu 4.99936, mengindikasikan tingkat kemiripan tertinggi dengan dokumen uji D100.

Perhitungan dengan P = 7

Hasil perhitungan kemiripan menggunakan Minkowski Distance dengan P = 7 dapat dilihat pada Tabel 8.

Tabel 8. Hasil Perhitungan Minkowski Distance dengan P = 7

Data Uji	Pembanding	Distance
D100	D1	4.423619
D100	D2	8.353126
D100	D3	11.15605
D100	D4	4.758393
D100	D5	7.071441
D100	D6	10.29613
D100	D7	9.517102
D100	D8	6.188354
D100	D9	7.825734
D100	D10	7.965309

Dari tabel di atas, hasil perhitungan menggunakan nilai P = 7 juga menunjukkan bahwa dokumen D1 memiliki jarak terkecil yaitu 4.423619, mengindikasikan tingkat kemiripan tertinggi dengan dokumen uji D100.

Implementasi Python

Implementasi metode Minkowski Distance dalam penelitian ini menggunakan bahasa pemrograman Python. Setelah melakukan perhitungan terhadap seluruh data, diperoleh bahwa dokumen D41 memiliki jarak Minkowski terkecil terhadap dokumen uji D100 ketika menggunakan nilai P = 7, dengan nilai jarak sebesar 3.838809.

Data	Parameter	Jarak	Rekomendasi
D41	P = 3	6.5808443652413 93	Jarak Lebih Besar dibandingkan nilai *p* yang lebih tinggi. Kurang sensitif terhadap perbedaan kecil antar dokumen.
D41	P = 5	4.376178290593927	Jarak lebih kecil dibandingkan *p* = 3. Sensitif namun tidak sebaik *p* = 7.
D41	P = 7	3.838808533456745	Jarak terkecil di antara semua nilai *p*. Sangat sensitif terhadap perbedaan kecil antar dokumen.

Gambar 3 Hasil Perhitungan

Pendekatan komputasi dengan algoritma berbasis jarak telah terbukti efektif dalam berbagai konteks analisis kemiripan teks (Hutapea & Silalahi, 2023) (Ahmad et al., 2021).

KESIMPULAN

Penelitian ini telah berhasil mengidentifikasi parameter optimal dalam metode Minkowski Distance untuk mendeteksi kemiripan abstrak skripsi, dengan nilai P=7 menunjukkan performa terbaik dalam mengukur kemiripan dokumen. Temuan ini memajukan bidang deteksi kemiripan teks dengan menyempurnakan sensitivitas pengukuran jarak melalui penyesuaian parameter, memberikan alternatif yang lebih fleksibel dibandingkan metode konvensional yang cenderung bersifat statis dalam pengukuran jarak. Kontribusi ilmiah dari penelitian ini terletak pada pemahaman bahwa sensitivitas yang tepat pada dimensi dengan perbedaan signifikan sangat penting dalam konteks perbandingan teks akademik, khususnya abstrak skripsi yang kaya akan terminologi dan struktur khusus. Implementasi metode ini dapat dikembangkan menjadi sistem otomatis untuk memverifikasi keaslian karya ilmiah di lingkungan akademik, sehingga mendukung integritas akademik dan mengurangi beban kerja manual dalam proses verifikasi. Untuk penelitian selanjutnya, perlu eksplorasi nilai P yang lebih tinggi atau implementasi pendekatan adaptif yang dapat menentukan nilai P optimal secara otomatis berdasarkan karakteristik dataset. Integrasi teknik deep learning dengan metode berbasis jarak juga menjanjikan untuk mengatasi keterbatasan pendekatan berbasis vektor kata dalam menangkap konteks semantik yang lebih kompleks pada dokumen akademik.

DAFTAR PUSTAKA

Ahmad, I., Borman, R. I., Caksana, G. G., & Fakhrurozi, J. (2021). Implementasi String Matching Dengan Algoritma Boyer-Moore Untuk Menentukan Tingkat Kemiripan Pada Pengajuan Judul Skripsi/Ta Mahasiswa (Studi Kasus: Universitas Xyz). *SINTECH (Science and Information Technology) Journal*, *4*(1), 53–58. https://doi.org/10.31598/sintechjournal.v4i1.699

Catur, W., Tulloh, R., & Wijayanti, D. E. (2023). *Identification of fingerprint image with Minkowski distance algorithm approach*. *3*(2), 69–78.

Euclidean, P. J., & Dalam, D. A. N. C. (2024). *Perbandingan jarak euclidean, cityblock, minkowski, canberra, dan chebyshev dalam sistem temu kembali citra batik*. *12*(3).

Hutapea, M. I., & Silalahi, A. P. (2023). Moderna's Vaccine Using the K-Nearest Neighbor (KNN) Method: An Analysis of Community Sentiment on Twitter. *Jurnal Penelitian Pendidikan IPA*, *9*(5), 3808–3814. https://doi.org/10.29303/jppipa.v9i5.3203

Kambey, G. E. I., Sengkey, R., & Jacobus, A. (2020). Penerapan Clustering pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia. *Jurnal Teknik Informatika*, *15*(2), 75–82.

Kurnia Aini, S. (2022). Perancangan Sistem Pendukung Keputusan Terhadap Pendeteksi Plagiarisme Judul Skripsi. *Teknologipintar.Org*, *2*(2), 1.

Kurniana, I. R., Muhima, R.R., Wardana, S., Hakimah, M. (2021). *Penerapan Algoritma K-Means Untuk Pengelompokan Topik Dokumen Studi Kasus:Dokumen Abstrak Skripsi Jurusan Teknik Informatika ITATS Kurniana,*. *1*, 219–224.

Lindang, D. N., Muniar, A. Y., Halid, A., Muhajirin, M., & Amiruddin, A. (2022). Sistem Penentuan Kemiripan Antar Skripsi Menggunakan Metode Cosine Similarity Pada Perpustakaan. *Sntei*, 321–324.

Risparyanto, A. (2020). Turnitin Sebagai Alat Deteksi Plagiarisme. *UNILIB: Jurnal Perpustakaan*, *11*(2), 126–135. https://doi.org/10.20885/unilib.vol11.iss2.art5

Vendyansyah, N., & Pranoto, Y. A. (2021). Perancangan dan Pembuatan Aplikasi untuk Mendeteksi Kemiripan Jawaban Menggunakan Cosine Similarity. *Jurnal Teknika (Jurnal Fakultas Teknik Universitas Islam Lamongan)*, *13*(1), 23–28.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Widaningrum, I., Mustikasari, D., Arifin, R., Tsaqila, S. L., & Fatmawati, D. (2022). Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) dan K-Means Clustering Untuk Menentukan Kategori Dokumen. *Prosiding Seminar Nasional Sistem Informasi Dan Teknologi (SISFOTEK)*, 145–149.