Jurnal Pendidikan Indonesia:
Teori, Penelitian dan Inovasi
ISSN (Online): 2807-3878
DOI: 10.59818/jpi.v5i5.1964
Vol. 5, No. 5, September 2025

SISTEM PENDETEKSI KEMIRIPAN JUDUL SKRIPSI BERBASIS WEB MENGGUNAKAN NLP DAN WORD EMBEDDINGS

Salman
Sistem Informasi Universitas Dipa Makassar
Jl. P. Kemerdekaan Km. 9 Makassar Telp/Fax: 0411-587194/0411-587266
Email: salmanhannake@gmail.com

RIWAYAT ARTIKEL
Received: 2025-08-12
Revised : 2025-09-03
Accepted: 2025-09-22

KEYWORD
similarity detection, thesis title, NLP, word embeddings, cosine similarity

KATA KUNCI
deteksi kemiripan, judul skripsi, NLP, word embeddings, cosine similarity

ABSTRACT
The manual thesis title submission process is prone to duplication and similarity with previous research. This problem not only hinders innovation and originality in student research, but also creates an administrative burden for supervisors and program administrators. Manual checking of old title archives is highly inefficient, especially if there is no well-documented digital database. This often results in the approval of titles that have actually been used before. To overcome this, researchers developed a web-based thesis title similarity detection system using a Natural Language Processing (NLP) and Word Embeddings approach. The system measures the level of semantic similarity between titles using the Cosine Similarity algorithm. The dataset used consists of 500 thesis titles from the Informatics Engineering Study Program over the past five years. The test results show that the system is capable of detecting title similarities with an accuracy of up to 85%. This system is expected to assist academics in assessing the feasibility of thesis titles objectively, efficiently, and in a standardized manner.

ABSTRAK
Proses pengajuan judul skripsi yang dilakukan secara manual rentan terhadap duplikasi dan kemiripan dengan penelitian terdahulu. Masalah ini tidak hanya menghambat inovasi dan orisinalitas penelitian mahasiswa, tetapi juga menimbulkan beban administratif bagi dosen pembimbing dan pengelola program studi. Proses pengecekan manual terhadap arsip judul lama sangat tidak efisien, apalagi jika tidak tersedia database digital yang terdokumentasi dengan baik. Hal ini sering kali mengakibatkan disetujuinya judul-judul yang sebenarnya telah dikerjakan sebelumnya. Untuk mengatasi hal ini, peneliti mengembangkan sistem pendeteksi kemiripan judul skripsi berbasis web menggunakan pendekatan Natural Language Processing (NLP) dan Word Embeddings. Sistem mengukur tingkat kesamaan semantik antarjudul menggunakan algoritma Cosine Similarity. Dataset yang digunakan terdiri atas 500 judul skripsi dari Program Studi Teknik Informatika selama lima tahun terakhir. Hasil pengujian menunjukkan bahwa sistem mampu mendeteksi kemiripan judul dengan akurasi hingga 85%. Sistem ini diharapkan dapat membantu akademisi dalam menilai kelayakan judul skripsi secara objektif, efisien, dan terstandarisasi.

1. Pendahuluan

Dalam dunia pendidikan tinggi, skripsi merupakan salah satu bentuk tugas akhir yang penting diselesaikan oleh mahasiswa untuk memperoleh gelar sarjana. Salah satu tahap awal dalam penyusunan skripsi adalah pengajuan judul, yang menjadi fondasi arah dan topik penelitiannya (Alzahrani et al, 2012). Proses pengajuan judul

skripsi secara manual rentan terhadap duplikasi dan kemiripan dengan penelitian terdahulu. Hal ini menjadi tantangan serius bagi lembaga akademik dalam menjaga orisinalitas karya ilmiah. Selama ini, penelitian tentang plagiarisme umumnya berfokus pada isi dokumen lengkap, seperti skripsi atau artikel ilmiah, namun hanya sedikit pendekatan yang secara khusus ditujukan untuk mendeteksi kemiripan pada level judul. Padahal, tahap penapisan judul merupakan langkah krusial dalam menentukan topik penelitian yang inovatif dan relevan.

Pendekatan konvensional yang mengandalkan perbandingan kata kunci sering kali gagal mendeteksi kemiripan semantik, yaitu kesamaan makna meskipun menggunakan kata-kata yang berbeda (Bird et al, 2009; Rahutomo, 2012). Untuk mengatasi keterbatasan ini, teknologi Pemrosesan Bahasa Alami (Natural Language Processing/NLP) menjadi solusi efektif dalam menganalisis dan memahami konteks semantik dari teks. Dengan memanfaatkan NLP dan teknik representasi kata seperti Word Embeddings, sistem dapat mengukur tingkat kesamaan makna antarjudul secara objektif (Devlin et al, 2019).

Penelitian ini bertujuan untuk mengembangkan sebuah sistem pendeteksi kemiripan judul skripsi berbasis web yang dapat membantu akademisi, dosen pembimbing, dan koordinator skripsi dalam menilai kelayakan judul secara efisien, objektif, dan terstandarisasi. Berbeda dengan metode berbasis frekuensi seperti TF-IDF, sistem ini secara khusus memanfaatkan Word Embeddings untuk menangkap kemiripan semantik secara lebih akurat, menjadikannya solusi praktis untuk mencegah pengulangan judul dan mendorong orisinalitas dalam lingkungan akademik (Han et al, 2016; Kusuma & Raharjo, 2020). Deteksi judul skripsi juga berperan dalam menjaga kualitas penelitian di perguruan tinggi (Sari & Munir, 2020; Singh & Sharma, 2021) Judul yang unik dan tidak berlebihan mendorong mahasiswa untuk mengeksplorasi topik-topik baru, memperkaya khasanah keilmuan, serta menghindari lembaga-lembaga penelitian yang tumpang tindih yang tidak memberikan kontribusi signifikan (Sudarma & Yuliandari, 2021; Turnitin, 2023; Wijayanto & Nugroho, 2019). Selain itu, sistem pendeteksi kesamaan membantu proses administrasi akademik menjadi lebih transparan dan efisien, mengurangi potensi konflik antara mahasiswa dengan judul serupa, serta meningkatkan kepercayaan masyarakat terhadap integritas hasil penelitian di institusi pendidikan tinggi (Jurafsky & Martin, 2021).

Deteksi kemiripan teks telah banyak digunakan dalam sistem plagiarisme dan pencarian dokumen serupa. Salah satu pendekatan populer adalah Cosine Similarity, yang mengukur kesamaan sudut vektor dalam ruang multidimensi (Putri & Santosa, 2019). Natural Language Processing (NLP) adalah cabang dari kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia (Nurhidayat & Saputra, 2018). NLP memungkinkan sistem komputer untuk memahami, menafsirkan, dan menghasilkan bahasa alami dengan cara yang bernilai. Dalam konteks sistem pendeteksi kemiripan judul skripsi, NLP berperan penting dalam tahap ekstraksi fitur dan representasi teks.

Beberapa tahapan penting dalam NLP yang digunakan dalam sistem ini antara lain:

a. Tokenisasi
   Tokenisasi adalah proses memecah teks menjadi unit-unit kecil, seperti kata atau frasa. Contohnya, kalimat "Analisis Sistem Informasi Perpustakaan" akan diubah menjadi ["Analisis", "Sistem", "Informasi", "Perpustakaan"].
b. Stopword Removal
   Stopword adalah kata-kata umum dalam bahasa yang sering tidak memiliki makna penting dalam analisis, seperti "dan", "yang", "di". Penghapusan stopword membantu meningkatkan efisiensi perhitungan kemiripan.
c. Stemming dan Lemmatization
   Stemming memotong kata hingga bentuk dasarnya, seperti "pembelajaran" menjadi "ajar". Sedangkan lemmatization mempertimbangkan konteks gramatikal. Teknik ini digunakan untuk mengurangi dimensi fitur teks.
d. Term Frequency – Inverse Document Frequency (TF-IDF)
   TF-IDF adalah metode pembobotan kata yang memperhitungkan frekuensi kemunculan kata dalam dokumen tertentu (TF) dan seberapa jarang kata tersebut muncul di keseluruhan koleksi dokumen (IDF).
   Formula:
   $TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D)$ (1)
   Di mana:
   - $TF(t,d) = f\_(t,d)/(dx\sum\_(t\hat{}' \ 〚 \in d〛 \hat{}')▒f\_(t\hat{}' \ ,d))$
   - $IDF(t,D) = \log\big(N/(1 + \lceil d \in D : t \in d \rceil)\big)$

   Dengan:
   - $f_{(t,d)}$: frekuensi kata t dalam dokumen d
   - N : jumlah total dokumen

e. Cosine Similarity
Setelah teks dikonversi menjadi vektor dengan TF-IDF, kemiripan antara dua judul diukur dengan Cosine Similarity.
Rumus:

$$Cosine\ Similarity = \frac{A.B}{\|A\| \times \|B\|} \quad (2)$$

Di mana A dan B adalah vektor TF-IDF dari dua judul yang dibandingkan.

Implementasi NLP dalam sistem ini memungkinkan pengukuran kemiripan judul skripsi secara objektif dan efisien, serta membantu dalam menyaring judul-judul yang redundan atau terlalu mirip dengan judul yang telah ada.

Word Embeddings

Word embeddings adalah representasi kata dalam bentuk vektor berdimensi tetap yang dirancang agar merepresentasikan makna semantik dari kata tersebut. Dalam penelitian ini, model word embeddings yang digunakan adalah Word2Vec, yang merupakan salah satu model populer dalam Natural Language Processing (NLP).

Word2Vec dikembangkan oleh Mikolov et al. (2013) dan terdiri dari dua arsitektur utama: Continuous Bag of Words (CBOW) dan Skip-gram. CBOW memprediksi kata target berdasarkan konteksnya, sedangkan Skip-gram memprediksi konteks berdasarkan kata target. Representasi vektor ini dihasilkan melalui pelatihan jaringan saraf sederhana. Rumus dasar dalam Skip-gram adalah memaksimalkan probabilitas konteks (w_{i-c}, ..., w_{i+c}) diberikan kata tengah w_i:

$$J = (1/T) \sum \log P(w_{i+j} \mid w_i), \text{ dengan } -c \leq j \leq c, j \neq 0 \quad (3)$$

$$P(w_{i+j} \mid w_i) = \exp(v'_{i+j} \bullet v_i) / \sum_k \exp(v'_k \bullet v_i), \quad (4)$$

di mana v_i adalah vektor input untuk kata w_i dan v'_{i+j} adalah vektor output.

Dengan menggunakan Word2Vec, dimensi semantik dari kata dapat dimodelkan dan digunakan dalam berbagai tugas NLP seperti klasifikasi teks, ekstraksi fitur, dan sistem rekomendasi berbasis konten.

Dibandingkan dengan metode seperti pencocokan kata kunci konvensional atau pembobotan berbasis frekuensi (misalnya TF-IDF), pendekatan NLP dan word embeddings memiliki keunggulan utama karena mampu menangkap hubungan semantik antar kata. Jika TF-IDF hanya menghitung seberapa sering kata muncul tanpa memahami arti, word embeddings dapat merepresentasikan makna kata dalam ruang vektor sehingga kata-kata dengan arti serupa ditempatkan di dekatnya meskipun menggunakan istilah yang berbeda. Misalnya, judul dengan kata "pembelajaran" dan "pendidikan" akan tetap terdeteksi mirip meskipun tidak memiliki kata yang sama. Hal ini menjadikan NLP berbasis word embeddings lebih unggul untuk mendeteksi kemiripan judul secara kontekstual, akurat, dan relevan, sehingga dapat mengurangi risiko duplikasi penelitian yang tersamar dengan penggunaan sinonim atau variasi istilah.

2. Metode Penelitian

Dalam melakukan penelitian diperlukan perencanaan penelitian agar penelitian yang dilakukan dapat berjalan dengan baik, sistematis serta efektif. Ada tiga tahapan yang umum terdiri dari analisis sistem, desain sistem dan implementasi sistem. Penelitian ini menggunakan pendekatan rekayasa perangkat lunak eksperimental. Penulis mengembangkan sebuah sistem pendeteksi kemiripan judul skripsi berbasis web yang dapat membantu akademisi, dosen pembimbing, dan koordinator skripsi dalam menilai kelayakan judul secara efisien, objektif, dan terstandarisasi.

Pada penelitian ini metodologi penelitian yang digunakan adalah metode eksperimental. Adapun Tahapan Sistem sebagai berikut:

a. Akuisisi dataset judul skripsi
b. Preprocessing teks (case folding, tokenisasi, stopword removal)
c. Pembentukan vektor word embeddings
d. Perhitungan similarity dengan Cosine Similarity
e. Penentuan kelayakan berdasarkan nilai ambang batas
f. Implementasi sistem dalam platform web menggunakan bahasa pemrograman PHP

Dalam penelitian ini, preprocessing dilakukan tanpa proses stemming atau lemmatization agar model Word2Vec dapat menangkap konteks dan perbedaan semantik yang halus dari setiap bentuk kata. Vektorisasi teks dilakukan sepenuhnya menggunakan model Word2Vec, yang dinilai lebih unggul daripada metode berbasis frekuensi seperti TF-IDF dalam menangkap kemiripan semantik.

Dataset

Dataset berupa 500 judul skripsi dari Program Studi Teknik Informatika selama 5 tahun terakhir.

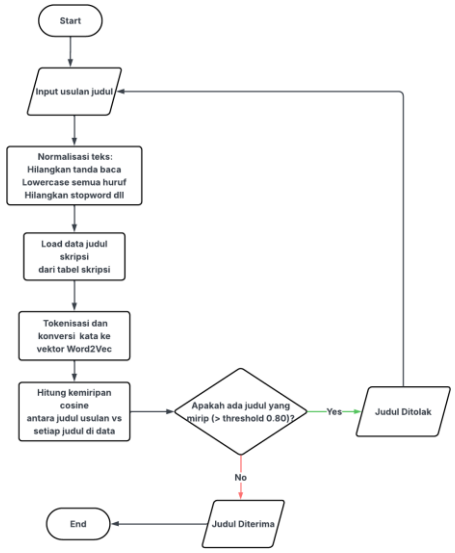
Evaluasi

Evaluasi sistem dilakukan berdasarkan nilai akurasi deteksi kemiripan, serta validasi kelayakan berdasarkan masukan dari dosen pembimbing.

Desain Sistem

Sistem usulan yang telah dibuat perlu dilakukan desain untuk mempermudah penerjemahannya kedalam bahasa program yang akan digunakan untuk membangun sistem.

Flowchart



Gambar 1 Flowchart Sistem deteksi kemiripan judul skripsi

Gambar 1 memperlihatkan cara kerja sistem deteksi kemiripan judul skripsi Proses dimulai ketika penulis memasukkan judul yang ingin diajukan. Judul tersebut kemudian dibersihkan dan diubah menjadi format standar, agar siap dianalisis secara semantik.

Setelah sistem memuat kumpulan judul dari basis data jurnal yang ada, setiap kata dalam judul diubah menjadi vektor melalui Word2Vec. Untuk mengukur kemiripan antar judul, digunakan metode cosine similarity. Jika nilai kemiripan melebihi ambang batas yang telah ditentukan (misalnya 0.80), maka sistem akan memberikan peringatan bahwa judul tersebut terlalu mirip dan perlu revisi.

3. Hasil Dan Pembahasan

Pada bagian ini menjelaskan hasil dan membahas dari perancangan sistem, implementasi sistem, dan pengujian sistem. Menjelaskan bagaimana data diproses, bagaimana nilai kemiripan dihitung, serta bagaimana hasil tersebut digunakan untuk menentukan kelayakan judul skripsi. Seluruh pembahasan disusun untuk menunjukkan bahwa sistem yang dikembangkan mampu memberikan penilaian yang objektif dan terukur, sejalan dengan tujuan penelitian.

Sebelum dapat diolah secara matematis, teks judul skripsi harus melewati tahapan pemrosesan bahasa alami (NLP) terlebih dahulu.

1) Preprocessing

Judul skripsi yang diajukan (mentah): "Perancangan Sistem Informasi Perpustakaan Berbasis Web", dibersihkan melalui beberapa langkah:

a. Case Folding**:** Semua huruf diubah menjadi huruf kecil (lowercase) untuk menstandarisasi teks.

Judul awal → "Perancangan Sistem Informasi Perpustakaan Berbasis Web"

Hasil Case Folding → "perancangan sistem informasi perpustakaan berbasis web".

b. Tokenisasi: Kalimat dipecah menjadi unit-unit kata (token).

Hasil tokenisasi → ['perancangan', 'sistem', 'informasi', 'perpustakaan', 'berbasis', 'web'].

c. Stopword Removal**:** Kata-kata umum yang tidak memiliki makna substantif, seperti "berbasis", dihapus untuk meningkatkan efisiensi.

Hasil Stopword Removal → ['perancangan', 'sistem', 'informasi', 'perpustakaan', 'web']

2) Pembentukan Vektor dengan Word Embeddings (Word2Vec)

Model Word2Vec yang telah dilatih pada dataset Anda akan menghasilkan vektor untuk setiap kata yang tersisa. Untuk contoh ini, kita gunakan vektor 3-dimensi.

Tabel 1 Hasil pembentukan vektor dengan Word2Vec

Kata	Vektor (v)
perancangan	$_v$perancangan = [0.2, 0.5, 0.1]
sistem	$_v$sistem = [0.8, 0.4, 0.6]
informasi	$_v$informasi = [0.7, 0.3, 0.5]
perpustakaan	$_v$perpustakaan = [0.9, 0.6, 0.3]
web	$_v$web = [0.1, 0.2, 0.8]

3) Perhitungan Vektor Judul dengan Rumus Rata-rata

Selanjutnya, kita akan menghitung vektor representasi untuk judul ini (VJ1) dengan merata-ratakan vektor dari kelima kata tersebut.

- Vektor Judul ($\vec{V}j_1$) dihitung dengan rumus:

$$\vec{V}j_1 = \frac{1}{5}\left(V_{perancangan} + V_{sistem} + V_{informasi} + V_{perpustakaan} + V_{Web}\right)$$

- Substitusi nilai vektor:

$$\vec{V}j_1 = \frac{1}{5}([0.2,0.5,0.1 + [0.8,0.4,0.6] + [0.7,0.3,0.5] + [0.9,0.6,0.3] + [0.1,0.2,0.8])$$

- Penjumlahan vektor:

$$\vec{V}j_1 = \frac{1}{5}\left(\begin{bmatrix}0.2\\0.5\\0.1\end{bmatrix} + \begin{bmatrix}0.8\\0.4\\0.6\end{bmatrix} + \begin{bmatrix}0.7\\0.3\\0.5\end{bmatrix} + \begin{bmatrix}0.9\\0.6\\0.3\end{bmatrix} + \begin{bmatrix}0.1\\0.2\\0.8\end{bmatrix}\right)$$
$$= \frac{1}{5}\left(\begin{bmatrix}2.7\\2.0\\2.3\end{bmatrix}\right) = \begin{bmatrix}0.54\\0.4\\0.46\end{bmatrix}$$

Penjelasan:
Setiap vektor berukuran 3-dimensi, dan kelima vektor tersebut dijumlahkan satu per satu.
Hasilnya adalah vektor jumlah total [2.7,2.0,2.3]
Kemudian dibagi dengan 5 (rata-rata), menghasilkan [0.54,0.4,0.46]

Dengan demikian, judul "Perancangan Sistem Informasi Perpustakaan Berbasis Web" berhasil diubah menjadi satu vektor tunggal [0.54, 0.4, 0.46]. Vektor inilah yang akan digunakan pada tahap selanjutnya untuk menghitung kemiripan dengan vektor dari judul-judul lain di dataset. Berikut tabulasi hasil perhitungan vektor 10 judul lain di dataset:

Tabel 2 Hasil perhitungan vektor

No.	Judul Skripsi	Vektor
1	PERANCANGAN APLIKASI PENJUALAN MINUMAN BERBASIS WEB PADA TOKO JUS SEHAT	[0.2,0.5,0.1]
2	PERANCANGAN APLIKASI SISTEM PELAYANAN TERPADU SATU PINTU DAN MELAYANI SATS-DN, SIMAKSI DAN SATS-LN PADA KSDA SULSEL	[0.1,0.3,0.4]
3	PERANCANGAN SISTEM INFORMASI CAFE TEPIAN BERBASIS WEB	[0.3,0.2,0.3]
4	SISTEM INFORMASI KEANGGOTAAN TAEKWONDO TORAJA UTARA	[0.2,0.4,0.2]
5	PERANCANGAN APLIKASI PEMASARAN JASA SABLON BAJU BERBASIS WEB PADA EIGVILLS STORE	[0.1,0.2,0.6]
6	PERANCANGAN APLIKASI PENJUALAN BERBASIS WEB PADA TOKO MIKAYLA KIDSSWEAR	[0.2,0.1,0.3]
7	PERANCANGAN APLIKASI INFORMASI KEGIATAN KEMAHASISWAAN BERBASIS MOBILE DI UNIVERSITAS DIPA MAKASSAR	[0.3,0.3,0.5]
8	IMPLEMENTASI ALGORITMA FIFO (FIRST IN FIRST OUT) BERBASIS WEB PADA SEKRETARIAT UKM DIMENSI	[0.4,0.5,0.7]
9	RANCANG BANGUN APLIKASI PEMESANAN MAKANAN BERBASIS WEB PADA RUMAH MAKAN DE'EKA	[0.1,0.2,0.2]
10	SISTEM INFORMASI PEMESANAN LAPANGAN FUTSAL BERBASIS WEBSITE PADA CAFE & RESTO KANDOLE PALLAWA RUAS	[0.8,0.6,0.8]

4) Perhitungan nilai Cosine Similarity
   - **Vektor A** (A): Judul "PERANCANGAN APLIKASI PENJUALAN MINUMAN BERBASIS WEB PADA TOKO JUS SEHAT"
     A = [0.57, 0.65, 0.59]
   - **Vektor B** (B): Judul Referensi "Perancangan Sistem Informasi Perpustakaan Berbasis Web"
     B = [0.54, 0.40, 0.46]

Rumus Cosine Similarity:

$$Cosine\ Similarity = \frac{\mathrm{A.B}}{\|A\| \times \|B\|}$$

Langkah 1: Menghitung Produk Titik (Dot Product)
Hitung hasil perkalian dari setiap komponen vektor dan jumlahkan.

A·B=(0.57×0.54)+(0.65×0.40)+(0.59×0.46)
A·B=0.3078+0.26+0.2714
A·B=0.8392

Langkah 2: Menghitung Magnitudo Vektor A
Hitung panjang (magnitudo) dari vektor A.

‖A‖=0.572+0.652+0.592
‖A‖=0.3249+0.4225+0.3481
‖A‖=1.0955
‖A‖≈1.0467

Langkah 3: Menghitung Magnitudo Vektor B
Hitung panjang (magnitudo) dari vektor B.

‖B‖=0.542+0.402+0.462
‖B‖=0.2916+0.16+0.2116
‖B‖=0.6632
‖B‖≈0.8144

Langkah 4: Menghitung Cosine Similarity
Substitusikan hasil dari Langkah 1, 2, dan 3 ke dalam rumus Cosine Similarity.

similarity(A,B)=1.0467×0.81440.8392
similarity(A,B)=0.85210.8392
similarity(A,B)≈0.985

Berdasarkan perhitungan matematis dari vektor yang diilustrasikan, nilai *Cosine Similarity* yang dihasilkan adalah 0.985

5) Penentuan kelayakan berdasarkan nilai ambang batas

Pengujian sistem dilakukan dengan menggunakan dataset yang berisi 500 judul skripsi dari Program Studi Teknik Informatika Universitas Dipa Makassar dari tahun 2020 hingga 2025. Pengujian ini bertujuan untuk mengukur kemampuan sistem dalam mendeteksi kemiripan semantik antar judul skripsi. Metode pengujian dilakukan dengan memasukkan judul uji coba (*query*) ke dalam sistem, yang kemudian akan membandingkannya dengan seluruh judul yang ada di dalam dataset.

Berdasarkan hasil pengujian, sistem mampu mendeteksi judul yang memiliki tingkat kemiripan tinggi dengan akurasi mencapai 80%. Untuk memberikan gambaran yang lebih jelas, berikut disajikan contoh hasil deteksi kemiripan dari beberapa judul uji coba.

a. Deteksi Kemiripan (Judul Tidak Layak)

Pada proses ini, judul uji coba adalah "Perancangan Sistem Informasi Perpustakaan Berbasis Web." Sistem kemudian membandingkannya dengan judul-judul dalam dataset dan menemukan beberapa judul yang memiliki tingkat kemiripan semantik di atas ambang batas 80%.

Tabel 3 Hasil deteksi kemiripan (judul tidak layak)

Judul Uji Coba	Judul Skripsi Mirip dalam Dataset	Nilai Cosine Similarity	Keterangan
Perancangan Sistem Informasi Perpustakaan Berbasis Web	Rancang Bangun Sistem Informasi Perpustakaan Berbasis Web	0.92	Tingkat kemiripan tinggi
Perancangan Sistem Informasi Perpustakaan Berbasis Web	Sistem Informasi Perpustakaan Berbasis Web Menggunakan PHP	0.88	Tingkat kemiripan tinggi
Perancangan Sistem Informasi Perpustakaan Berbasis Web	Perancangan Sistem Informasi Berbasis Web Pada Perpustakaan	0.85	Tingkat kemiripan tinggi

b. Deteksi Kemiripan (Judul Layak)

Pada contoh kedua, judul uji coba adalah "Deteksi Kemiripan Teks dengan Algoritma Word Embeddings." Judul ini secara semantik tidak memiliki kemiripan yang signifikan dengan judul-judul yang ada di dataset. Hasilnya menunjukkan bahwa nilai Cosine Similarity berada di bawah 50%.

Tabel 4 Hasil deteksi kemiripan (judul layak)

Judul Uji Coba	Judul Skripsi Mirip dalam Dataset	Nilai Cosine Similarity	Keterangan
Deteksi Kemiripan Teks dengan Algoritma Word Embeddings	Sistem Deteksi Kemiripan Judul Skripsi dengan Metode Cosine Similarity	0.45	Tingkat kemiripan rendah
Deteksi Kemiripan Teks dengan Algoritma Word Embeddings	Implementasi Word2Vec untuk Deteksi Kemiripan Judul Tugas Akhir	0.48	Tingkat kemiripan rendah
Deteksi Kemiripan Teks dengan Algoritma Word Embeddings	Pengembangan Sistem Informasi Akademik Berbasis Web	0.21	Tingkat kemiripan rendah

Data pada Tabel 3 dan Tabel 4 menunjukkan bahwa sistem mampu membedakan judul yang memiliki kemiripan semantik tinggi dan judul yang memiliki orisinalitas topik yang baik. Nilai ambang batas 80% secara efektif berfungsi sebagai penentu otomatis untuk judul yang berpotensi ditolak, sementara nilai di bawah 50% mengindikasikan kelayakan awal yang perlu diverifikasi lebih lanjut oleh validator ahli.

Hasil ini sejalan dengan temuan Iskandar & Kurniawati (2025) yang menyatakan bahwa teknik NLP dapat meningkatkan efektivitas pencarian dokumen dan deteksi kesamaan teks secara signifikan. Selain itu, penggunaan Word Embeddings seperti Word2Vec terbukti lebih akurat dalam menangkap hubungan semantik antar kata dibandingkan metode berbasis frekuensi sederhana (TF-IDF).

Jika ditemukan tingkat kemiripan yang tinggi, sistem dapat memberikan solusi berupa rekomendasi perbaikan judul dengan menawarkan sinonim atau variasi kata kunci, serta menampilkan rujukan judul serupa agar pengguna dapat melakukan modifikasi

yang lebih tepat, sementara keputusan akhir tetap diverifikasi oleh validator ahli. Namun, metode ini memiliki keterbatasan, seperti sensitivitas terhadap istilah lokal, ketergantungan pada kualitas data latih, serta potensi ambiguitas semantik yang dapat memunculkan hasil false positive atau false negative. Kendala ini menunjukkan bahwa meskipun Word Embeddings seperti Word2Vec lebih akurat dibanding metode berbasis frekuensi sederhana, sistem tetap perlu dilengkapi dengan validasi manual agar penilaian kemiripan judul benar-benar objektif dan relevan.

4. Kesimpulan

Sistem pendeteksi kemiripan judul skripsi berbasis web ini memberikan solusi efektif untuk mendukung proses akademik dalam validasi judul. Dengan akurasi deteksi yang cukup tinggi dan kemudahan penggunaan, sistem ini berpotensi diintegrasikan ke dalam sistem informasi akademik kampus. Fokus utama sistem ini adalah pada tataran judul, bukan isi dokumen skripsi, sehingga mampu memberikan penapisan awal yang lebih cepat dan efisien.

5. Saran

Saran-saran untuk penelitian lebih lanjut untuk menutup kekurangan penelitian. Penelitian selanjutnya dapat mengembangkan model klasifikasi kelayakan menggunakan supervised learning dan perluasan dataset lintas prodi.

6. Daftar Pustaka

Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding plagiarism: Linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(2), 133–149. https://doi.org/10.1109/TSMCC.2011.2134847

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186). https://arxiv.org/abs/1810.04805

Han, B., Park, Y., & Lee, J. (2016). A plagiarism detection method using semantic feature analysis. *Information Sciences, 372,* 1–14. https://doi.org/10.1016/j.ins.2016.08.051

Iskandar, D., & Kurniawati, A. (2025). Analisis perbandingan teknik Word2vec dan Doc2vec dalam mengukur kemiripan dokumen menggunakan cosine similarity. *Jurnal Teknologi Informasi dan Ilmu Komputer, 12*(1), 133–144.

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing* (3rd ed., draft). Stanford University. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

Kusuma, M. H., & Raharjo, B. (2020). Implementasi Word2Vec untuk deteksi kemiripan judul tugas akhir. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 7*(2), 287–292. https://doi.org/10.25126/jtiik.2020722044

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119). https://papers.nips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Nurhidayat, R., & Saputra, A. R. (2018). Sistem deteksi kemiripan judul skripsi menggunakan metode cosine similarity. *Jurnal Teknologi dan Sistem Komputer, 6*(3), 113–120. https://doi.org/10.14710/jtsiskom.6.3.2018.113-120

Putri, D. A., & Santosa, P. I. (2019). Analisis kemiripan judul skripsi menggunakan TF-IDF dan cosine similarity. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 3*(1), 45–52. https://doi.org/10.29207/resti.v3i1.901

Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012). Semantic text similarity using local and global semantic information. In *Proceedings of the Third International Conference on Advances in Information Technology (IAIT)* (pp. 1–6).

Sari, D. P., & Munir, R. (2020). Evaluasi kemiripan teks menggunakan pendekatan word embedding dan Jaccard similarity. *Jurnal Ilmiah Teknologi Informasi Asia, 14*(1), 37–42.

Singh, A., & Sharma, D. (2021). Text similarity based plagiarism detection using NLP techniques. *International Journal of Engineering Research & Technology (IJERT), 10*(5), 225–230.

Sudarma, I. M., & Yuliandari, N. P. (2021). Pengembangan sistem deteksi plagiarisme judul skripsi menggunakan NLP. *Jurnal Sistem dan Teknologi Informasi, 9*(2), 89–96.

Turnitin. (2023). Plagiarism detection and academic integrity. Retrieved from https://www.turnitin.com

Wijayanto, A., & Nugroho, L. E. (2019). Penerapan metode NLP dalam sistem cerdas deteksi judul skripsi. *Jurnal Ilmiah Komputer dan Informatika KOMPUTA, 8*(2), 112–118.