



## Deteksi Tingkat Kemiripan Judul Skripsi Prodi Informatika menggunakan Metode *Support Vector Machine & Natural Language Processing*

<u>Hanjas Haruna</u>	<u>INFO ARTIKEL</u>
Hanjas Haruna Universitas Muhammadiyah Makassar <a href="mailto:hanjasharuna24@gmail.com">hanjasharuna24@gmail.com</a>	ISSN: 3026-3603 Vol. 2, No. 1 April 2024 <a href="http://jurnal.ardenjaya.com/index.php/ajst">http://jurnal.ardenjaya.com/index.php/ajst</a>
Rizki Yusliana Bakti Universitas Muhammadiyah Makassar	
Titin Wahyuni Universitas Muhammadiyah Makassar	

© 2024 Arden Jaya Publisher All rights reserved

### **Saran Penulisan Referensi:**

Haruna, H., Bakti, R. Y., & Wahyuni, T. (2024). Deteksi Tingkat Kemiripan Judul Skripsi Prodi Informatika menggunakan Metode Support Vector Machine & Natural Language Processing. *Arus Jurnal Sains dan Teknologi*, 2 (2), 447-452.

### **Abstrak**

Skripsi merupakan salah satu persyaratan akhir dalam pendidikan akademis mahasiswa yang harus ditulis sesuai dengan bidang ilmunya. Proses awal dalam pembuatan skripsi adalah pengajuan judul, di mana mahasiswa harus memastikan bahwa judul yang diajukan tidak mirip dengan judul skripsi sebelumnya untuk menghindari plagiarisme. Deteksi kemiripan judul sangat penting untuk mencegah plagiarisme dan meningkatkan akurasi sistem rekomendasi dengan mengidentifikasi judul yang relevan. Dengan kemajuan teknologi informasi, tantangan seperti plagiarisme menjadi lebih kompleks, memerlukan sistem yang efisien untuk mengecek kemiripan judul. Salah satu metode yang digunakan adalah Support Vector Machine (SVM), sebuah teknik pembelajaran mesin yang efektif dalam klasifikasi teks, termasuk judul skripsi, berkat kemampuannya dalam menangani hubungan non-linear melalui kernel. Penggunaan Support vector machine dalam deteksi kemiripan judul bertujuan untuk memastikan keakuratan dan efisiensi dalam mengidentifikasi potensi plagiarisme. Kombinasi metode Support Vector Machine (SVM) dan Natural Language Processing (NLP), khususnya melalui vektorisasi menggunakan TF-IDF, telah terbukti efektif dalam mendeteksi kemiripan judul skripsi. Proses ini melibatkan pengumpulan data dari database judul\_skripsi, diikuti dengan preprocessing yang meliputi tokenisasi, penghapusan kata-kata tidak relevan, dan stemming. Data yang bersih dan terstruktur ini kemudian diubah menjadi representasi numerik melalui TF-IDF, yang memberikan bobot lebih tinggi pada kata-kata yang spesifik namun signifikan. Model SVM yang dilatih dengan data ini berhasil mendeteksi kemiripan dengan tingkat akurasi yang memadai.

**Kata kunci :** Deteksi Kemiripan, (SVM), (NLP), Klasifikasi Teks, Pembelajaran Mesin

### Abstract

A thesis is one of the final requirements in a student's academic education that must be written in accordance with their field of study. The initial process in writing a thesis is submitting a title, where students must ensure that the title submitted is not similar to the title of a previous thesis to avoid plagiarism. Detecting title similarity is very important to prevent plagiarism and improve the accuracy of the recommendation system by identifying relevant titles. With the advancement of information technology, challenges such as plagiarism have become more complex, requiring an efficient system to check title similarity. One of the methods used is the Support Vector Machine (SVM), a machine learning technique that is effective in text classification, including thesis titles, thanks to its ability to handle non-linear relationships through kernels. The use of Support Vector Machines in title similarity detection aims to ensure accuracy and efficiency in identifying potential plagiarism. The combination of the Support Vector Machine (SVM) and Natural Language Processing (NLP) methods, especially through vectorization using TF-IDF, has proven effective in detecting the similarity of thesis titles. This process involves collecting data from the thesis\_title database, followed by preprocessing including tokenization, removal of irrelevant words, and stemming. This clean and structured data is then transformed into a numeric representation through TF-IDF, which gives higher weight to specific but significant words. The SVM model trained with this data successfully detects similarities with a reasonable level of accuracy.

**Keywords:** Similarity Detection, Thesis Title, Informatics Study Program. (SVM), (NLP), (NLP), Text Classification, Machine Learning.

## A. Pendahuluan

Skripsi merupakan salah satu persyaratan akhir pendidikan akademis mahasiswa yang ditulis dalam bentuk karya tulis ilmiah serta sesuai bidang ilmunya. Tahapan pertama dalam prosedur skripsi adalah proses pengajuan judul. Mahasiswa wajib mengajukan judul skripsi dengan tingkat kemiripan yang tidak sama dengan judul skripsi sebelumnya agar tidak terjadi plagiarisme judul skripsi (Desena & Solichin, 2021)

Deteksi tingkat kemiripan judul adalah proses untuk mengukur seberapa mirip dua atau lebih judul dalam konteks tertentu, seperti penelitian akademis. Pentingnya deteksi kemiripan judul yaitu sebagai pengecekan plagiarisme untuk membantu mengidentifikasi judul-judul yang mungkin menjiplak atau sangat mirip dengan karya yang lain serta meningkatkan akurasi sistem rekomendasi dengan menemukan judul yang relevan dan menarik bagi pengguna (Fitrianingsih et al., 2022)

Dengan adanya perkembangan teknologi informasi saat ini membuat pertukaran informasi semakin mudah tidak hanya membawa dampak positif saja. Dampak negatif dari penggunaan pun tidak sedikit, salah satunya sebagai sarana untuk melakukan tindakan plagiarisme atau kegiatan mencuri kegiatan hasil karya orang lain dan mengakuinya sebagai karya sendiri dalam proses penyelesaian tugas akhir sehingga diperlukannya sebuah sistem untuk melakukan pengecekan plagiarisme pada judul ataupun tugas akhir mahasiswa dengan cepat dan tepat menggunakan yang digunakan salah satunya adalah Support Vector Machine (SVM) (Putra Bayu Pratama et al., 2021)

Support Vector Machine (SVM) adalah metode pembelajaran mesin yang efektif untuk klasifikasi teks, termasuk judul penelitian. Support Vector Machine (SVM) bekerja dengan baik pada dataset yang terstruktur dan memiliki kemampuan menangani hubungan non-linear melalui kernel trik. Penggunaan Support Vector Machine (SVM) dalam mendeteksi judul dapat mengotomatisasi klasifikasi dan memberikan hasil yang akurat, membantu akademisi dan peneliti dalam mengevaluasi dan memahami perkembangan dibidang teknik Informatika. (Desena & Solichin, 2021)

Natural Language Processing adalah cabang dari kecerdasan buatan (AI) yang berfokus pada interaksi antara komputer dan bahasa manusia. Natural Language Processing (NLP) memungkinkan komputer untuk membaca, memahami dan menghasilkan teks atau ucapan manusia dalam cara yang bermanfaat. (Fitrianingsih et al., 2022)

## B. Metodologi

### Teknik Pengujian Sistem

Pengujian sistem akan dilakukan untuk mengetahui apakah sistem yang dibangun sudah sesuai dengan analisa dan tujuan dari dibangunnya sistem ini. Untuk mengetahui hasil dari sistem ini apakah sudah sesuai dengan yang diharapkan, dilakukan penghitungan similarity dengan inputan judul yang diuji.

Rencana Pengujian Dalam tahapan rencana pengujian akan dilakukan sebagai berikut:

- a. Untuk tahapan pengujian sistem akan dilakukan menggunakan 10 judul yang akan dibandingkan dengan 760 judul yang tersimpan pada Database jurnal yang digunakan jurnal ISSN yang di dapat dari google scholar pada Universitas Budi Luhur. (Desena & Solichin, 2021)
- b. Dalam tahapan pengujian pertama dilakukan dengan cara membandingkan judul penelitian yang sama dengan judul penelitian yang tersimpan dalam database maka terlihat tingkat kemiripannya bobot cosine similarity berapa persen. (Mawanta et al., 2021)
- c. Dalam tahapan pengujian kedua inputan awal dengan menggunakan judul yang sama dengan konsep mengubah susunan kata atau kombinasi acak kata dalam judul penelitian, maka tingkat nilai cosine similarity hasil akan dibandingkan dengan judul tanpa mengubah susunan kata apakah bobot nilai cosine similarity akan berubah apa akan tetap sama. (Sheva Ricko Apriansyah, 2023)

Dalam pengujian ketiga dilakukan dengan cara penghapusan tiga kata di judul yang sama maka akan dilihat nilai tingkat kemiripan cosine similarity berapa persen apakah nilai cosine similarity apakah akan turun bobot kemiripannya. (Lestari & Sumarlinda, 2023)

### Teknik Analisis Data

Analisis data adalah proses pencarian dan pengaturan sistematika hasil data ulasan dan bahan-bahan yang dikumpulkan untuk meningkatkan pemahaman terhadap semua hal yang dikumpulkan dan memungkinkan menyajikan apa yang ditemukan. Analisis data yang digunakan dalam penelitian adalah melalui tahapan sebagai berikut :

- a. Pengumpulan Data  
Kumpulkan dataset yang berisi judul-judul program studi informatika dari berbagai institusi. Pastikan data sudah di-label jika ada informasi tentang tingkat kemiripan untuk digunakan sebagai data pelatihan. (Setiawan & Saprudin, 2023)
- b. Pembangunan Model  
Gunakan Support Vector Machine (SVM) sebagai algoritma klasifikasi, bagilah data menjadi set pelatihan dan set pengujian untuk evaluasi model dan Latih model metode Support Vector Machine (SVM) menggunakan data pelatihan yang telah diproses. Gunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score untuk menilai performa model. (Ahmad et al., 2020)
- c. Deteksi Kemiripan  
Terapkan model terlatih untuk mengukur kemiripan judul-judul baru, Gunakan threshold tertentu pada skor probabilitas dari model metode Support Vector Machine (SVM) untuk menentukan apakah dua judul dianggap mirip atau tidak. (Sanjaya et al., 2016)
- d. Interpretasi Hasil  
Visualisasikan hasil kemiripan dengan menggunakan diagram atau heatmap untuk memperlihatkan pasangan judul yang mirip. Interpretasikan hasil untuk memahami pola kemiripan yang ditemukan oleh model. (Hasanah et al., 2020)

## C. Hasil dan Pembahasan

Pada bagian ini akan di bahas hasil penelitian Data dikumpulkan dari database judul\_skripsi, yang terdiri dari nomor dan judul, dengan total 760 judul skripsi. Proses ini melibatkan ekstraksi data dari sistem basis data MySQL, memastikan bahwa setiap entri berisi informasi yang relevan dan akurat untuk keperluan analisis lebih lanjut. Data ini nantinya akan digunakan sebagai input dalam berbagai tahap preprocessing dan pelatihan model untuk mendeteksi kemiripan judul skripsi.

Saved to this PC

Tabel 1. Pengumpulan Data

No	Judul Skripsi
1	DESAIN DAN IMPLEMENTASI TIME BASED ONE TIME PASSWORD UNTUK MENINGKATKAN KEAMANAN OTENTIKASI PADA WEBSITE TOP UP GAME
2	IMPLEMENTASI AUGMENTED REALITY PADA PEMBUATAN GAME MOBILE DALAM MEMPERKENALKAN SEJARAH KEMERDEKAAN REPUBLIK INDONESIA
3	PEMBUATAN VERIFIKASI SERTIFIKAT DIGITALL SEBAGAI BUKTI KEABSAHAN MENGGUNAKAN ALGORITMA STEGANOGRAFI DENGAN METODE LEAST SIGNIFICANT BIT INSERTTION (LSR)
4	PENERAPAN FIKHI SHALAT DENGAN METODE TARJIH PADA APLIKASI MOBILE DAN ANIMASI 3D
5	PENERAPAN WEB SEMANTIK UNTUK PENCARIAN JUDUL SKRIPSI PADA RUANG BACA FAKULTAS TEKNIK UNIVERSITAS MUHAMMADIYAH MAKASSAR
6	PENCARIAN BUKU YANG RELEVAN DENGAN JUDUL SKRIPSI PADA RUANG BACA FAKULTAS TEKNIK UNISMUH MENGGUNAKAN ALGORITMA NAÏVE BAYES CLASSIFIER

memungkinkan pengguna untuk memasukkan judul skripsi baru dan melihat seberapa mirip judul tersebut dengan judul-judul skripsi yang sudah ada. Aplikasi akan memberikan skor yang menunjukkan seberapa mirip judul baru tersebut dengan yang ada dalam database.

Gambar 1. Tampilan Fitur deteksi kemiripan judul

### Deteksi Kemiripan Judul Skripsi

Masukkan Judul Skripsi:

Deteksi

Berdasarkan hasil yang ditampilkan, judul skripsi yang diusulkan "ANALISIS SENTIMEN APLIKASI GOPAY BERDASARKAN ULASAN DI PLAY STORE MENGGUNAKAN METODE NAIVE BAYES" memiliki kemiripan yang sangat tinggi dengan judul skripsi yang ada di database. Skor kemiripan teks adalah 92.93%, menunjukkan bahwa kedua judul memiliki kesamaan kata-kata yang identik. Rata-rata panjang string adalah 13 karakter, mengindikasikan konsistensi dalam panjang kata-kata. Persentase kemiripan hampir sama yang berarti kedua judul tersebut tidak hanya memiliki kata-kata yang sama, tetapi juga memiliki struktur yang sama.

**Deteksi Kemiripan Judul Skripsi**

Masukkan Judul Skripsi:

ANALISIS SENTIMEN APLIKASI GOPAY BERDASARKAN ULASAN DI PLAY STORE MENGGUNAKAN METODE NAIVE BAYES

Deteksi

Hasil Deteksi:

No	Judul Usulan	Judul Skripsi	Similarity Text	Rata-rata Panjang String	Persentase Kemiripan
1	ANALISIS SENTIMEN APLIKASI GOPAY BERDASARKAN ULASAN DI PLAY STORE MENGGUNAKAN METODE NAIVE BAYES	ANALISIS SENTIMEN APLIKASI REKSADANA BERDASARKAN ULASAN DI PLAY STORE MENGGUNAKAN METODE NAIVE BAYES	92.93	13	92.93%

Ajukan Judul

**Gambar 2.** Deteksi Kemiripan Judul Skripsi

#### D. Kesimpulan

1. Kombinasi metode Support Vector Machine (SVM) dan Natural Language Processing (NLP), khususnya melalui vektorisasi menggunakan TF-IDF, telah terbukti efektif dalam mendeteksi kemiripan judul skripsi. Proses ini melibatkan pengumpulan data dari database judul\_skripsi, diikuti dengan preprocessing yang meliputi tokenisasi, penghapusan kata-kata tidak relevan, dan stemming. Data yang bersih dan terstruktur ini kemudian diubah menjadi representasi numerik melalui TF-IDF, yang memberikan bobot lebih tinggi pada kata-kata yang spesifik namun signifikan. Model SVM yang dilatih dengan data ini berhasil mendeteksi kemiripan dengan tingkat akurasi yang memadai.
2. Jumlah data yang digunakan memiliki peran krusial dalam meningkatkan akurasi model dalam mendeteksi kemiripan judul. Semakin besar jumlah data yang tersedia, semakin baik model dapat mengenali pola-pola tertentu dalam penulisan judul, sehingga memperkuat kemampuan aplikasi dalam memberikan hasil yang lebih akurat dan andal di berbagai konteks penggunaan.
3. Hasil evaluasi menunjukkan akurasi sebesar 0.7982456140350878, presisi sebesar 0.8260869565217391, recall sebesar 0.7851239669421488, dan F1-Score sebesar 0.8050847457627118. Judul yang paling mirip dengan judul yang diajukan adalah "Sistem Informasi Perpustakaan" dengan skor kemiripan sebesar 0.4789599606584551. Perhitungan manual kemiripan kosinus juga menunjukkan berapa besar kemiripannya.

#### Saran

Memperluas cakupan data judul skripsi dari berbagai disiplin ilmu untuk lebih mengenali berbagai pola dan variasi dalam penulisan judul sehingga meningkatkan akurasi deteksi kemiripan.

#### E. Referensi

- Ahmad, I., Borman, R. I., Fakhrurozi, J., & Caksana, G. G. (2020). Software Development Dengan Extreme Programming (XP) Pada Aplikasi Deteksi Kemiripan Judul Skripsi Berbasis Android. *INOVTEK Polbeng-Seri Informatika*, 5(2), 297. <https://doi.org/10.35314/isi.v5i2.1654>
- Alamsyah, N., & Rasyidan, M. (2019). Deteksi Plagiarisme Tingkat Kemiripan Judul Skripsi Pada Fakultas Teknologi Informasi Menggunakan Algoritma Winnowing. *Technologia: Jurnal Ilmiah*, 10(4), 197. <https://doi.org/10.31602/tji.v10i4.2361>
- Desena, W., & Solichin, A. (2021). Pencarian Abstrak Tugas Akhir Mahasiswa Berdasarkan Tingkat Kemiripan Menggunakan Algoritma Winnowing dan Jaccard Similarity pada Universitas Budi Luhur. *Informatik: Jurnal Ilmu Komputer*, 17(2), 112. <https://doi.org/10.52958/iftk.v17i2.3628>

- Dillak, R. Y., Laumal, F., & Kadja, L. J. (2016). Sistem Deteksi Dini Plagiarisme Tugas Akhir Mahasiswa Menggunakan Algoritma Ngrams Dan Winnowing. *Jurnal Ilmiah Flash*, 2(1), 12. <https://doi.org/10.32511/jiflash.v2i1.19>
- Fitrianiingsih, N., Asfi, M., Prasetyo, D., Kusuma, R. P., & Sulhan, M. A. (2022). Deteksi Tingkat Kemiripan Judul Menggunakan Algoritma Oliver Pada Sistem Informasi Pengajuan Skripsi. *Jurnal Media Informatika Budidarma*, 6(4), 2148. <https://doi.org/10.30865/mib.v6i4.4409>
- Hasanah, U. N., Satra, R., & Umar, F. (2020). Deteksi Kemiripan Judul Skripsi Menggunakan Algoritma Smith Waterman. *Buletin Sistem Informasi Dan Teknologi Islam*, 1(1), 56–65. <https://doi.org/10.33096/busiti.v1i1.676>
- Islamiyati, D. S., Fikri, A., Teknik, F., Informatika, T., & Pamulang, U. (2022). Penerapan Algoritma Knuth-Morris-Pratt dalam Mendeteksi Tingkat Kemiripan Judul Skripsi Berbasis Web. 3(2), 58–62. <https://doi.org/10.47065/josh.v3i2.1168>
- Lestari, W., & Sumarlinda, S. (2023). Studi Komparatif Model Klasifikasi Kerentanan Penyakit Jantung Menggunakan Algoritma Machine Learning. *SATIN - Sains Dan Teknologi Informasi*, 9(1), 107–115. <https://doi.org/10.33372/stn.v9i1.918>
- Mawanta, I., Gunawan, T. S., & Wanayumini, W. (2021). Uji Kemiripan Kalimat Judul Tugas Akhir dengan Metode Cosine Similarity dan Pembobotan TF-IDF. *Jurnal Media Informatika Budidarma*, 5(2), 726. <https://doi.org/10.30865/mib.v5i2.2935>
- Processing, N. L., & Matching, F. S. (2022). *Natural language processing*.
- Sanjaya, A., Fauzi, I., & Uddin, M. F. (2016). Pencegah plagiasi dengan deteksi kemiripan judul skripsi. *Nusantara Oof Engineering*, 3(2), 7–11.
- Setiawan, A., & Saprudin. (2023). Implementasi Sistem Pendeteksian Kemiripan Judul Skripsi Dengan Algoritma Levenshtein Distance Pada Perpustakaan Universitas Pamulang. *Jurnal Ilmu Komputer Dan Pendidikan*, 2(1), 1–5. <https://journal.mediapublikasi.id/index.php/logic>
- Sheva Ricko Apriansyah. (2023). Penerapan Algoritma Winnowing Untuk Mendeteksi Kemiripan Judul Skripsi Pada Program Studi Informatika Universitas Teknologi Sumbawa. 1(2), 1–120.
- Sugiarto, L., Mulyadi, C., & Rihastuti, S. (2021). Analisa Algoritma String Matching Dan Winnowing Untuk Deteksi Kemiripan Judul Tugas Akhir Perguruan Tinggi. *Jurnal Teknologi Informasi*, 6(2), 97–106. <https://doi.org/10.52643/jti.v6i2.1141>
- Wahyuni, M. S., Setiawan, D., & Syahputra, T. (2021). Sistem Temu Kembali Informasi Dengan Latent Semantic Analisis Pada Kesamaan Tugas Akhir Mahasiswa. *Jurnal Teknisi*, 1(2), 46. <https://doi.org/10.54314/teknisi.v1i2.655>