

## ABSTRACT

### OPTIMIZATION OF JARO-WINKLER AND LEVENSHTTEIN FOR A THESIS TITLE FILTERING SYSTEM IN THE UMB FACULTY OF ENGINEERING

Name : Muhammad Adriansyah

Student ID : 1955201156

Supervisor : Dedy Abdullah, S.T., M.Eng.

*is crucial to prevent duplication of scientific work in the Faculty of Engineering, Universitas Muhammadiyah Bengkulu.<sup>2</sup> This study examines the optimization of the similarity threshold for Jaro-Winkler and Levenshtein Distance algorithms using 50 dummy thesis title pairs.<sup>3</sup> Binary evaluation shows that the Jaro-Winkler algorithm achieves optimal performance with an F1-score of 1.000 at a 0.85 threshold. Meanwhile, Levenshtein Distance achieves its optimal performance at a 0.70 threshold with an F1-score of 0.837. The Wilcoxon Signed-Rank test confirms a statistically significant difference in performance between the two algorithms with The title screening system is recommended to implement the Jaro-Winkler algorithm.*

**Keywords:** Jaro-Winkler, Levenshtein, Threshold, Thesis.

## ABSTRAK

### OPTIMALISASI JARO-WINKER DAN LEVENSHTTEIN UNTUK SISTEM PENYARINGAN JUDUL SKRIPSI DI FAKULTAS TEKNIK UMB

Nama : Muhammad Adriansyah

NPM : 1955201156

Pembimbing : Dedy Abdullah, S.T., M.Eng.

Penyaringan judul skripsi penting dilakukan untuk mencegah duplikasi karya ilmiah di Fakultas Teknik Universitas Muhammadiyah Bengkulu. Penelitian ini menguji optimasi batas ambang kemiripan (*threshold similarity*) pada algoritma Jaro-Winkler dan Levenshtein Distance menggunakan 50 pasangan judul skripsi fiktif.<sup>3</sup> Pengujian biner menunjukkan bahwa algoritma Jaro-Winkler mencapai kinerja optimal dengan *F1-score* sebesar 1,000 pada *threshold* 0,85. Sementara itu, Levenshtein Distance mencapai performa optimal pada *threshold* 0,70 dengan *F1-score* sebesar 0,837. Hasil uji Wilcoxon Signed-Rank membuktikan adanya perbedaan performa yang signifikan secara statistik antara kedua algoritma dengan nilai.

**Kata kunci : Treshold similiarity, Laveinshtein, Jaro-Winkler.**

# **BAB I**

## **PENDAHULUAN**

### **1.1 Latar Belakang**

Dalam lingkungan akademik, terutama pada jenjang pendidikan tinggi, keberadaan sistem pengelolaan skripsi yang efektif dan efisien merupakan aspek penting untuk menjamin orisinalitas karya ilmiah mahasiswa. Berdasarkan data Direktorat Jenderal Pendidikan Tinggi (DIKTI, 2023), kasus kemiripan atau duplikasi judul tugas akhir meningkat seiring dengan pertumbuhan jumlah mahasiswa dan keterbatasan sistem pengawasan digital di kampus. Permasalahan ini tidak hanya menimbulkan hambatan administratif, tetapi juga berpotensi menurunkan kredibilitas dan kualitas penelitian apabila tidak terdeteksi secara dini (Rahman et al., 2022)

Penelitian ini berfokus pada optimasi threshold similarity untuk kedua algoritma tersebut agar dapat mencapai keseimbangan terbaik antara precision dan recall dalam konteks sistem penyaringan judul skripsi. Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan sistem deteksi kemiripan teks yang lebih adaptif dan akurat, khususnya di lingkungan perguruan tinggi. Dengan adanya optimasi threshold yang tepat, sistem penyaringan judul skripsi dapat membantu pihak administrasi dan dosen pembimbing dalam meminimalisasi kemiripan judul serta meningkatkan efisiensi proses pengajuan proposal penelitian. Berdasarkan latar belakang tersebut, penelitian ini berupaya

menjawab beberapa permasalahan utama. Pertama, bagaimana kinerja algoritma Jaro–Winkler dan Levenshtein dalam mendeteksi tingkat kemiripan judul skripsi di lingkungan FT UMB. Kedua, bagaimana pengaruh variasi nilai *threshold similarity* terhadap tingkat akurasi sistem dalam mengklasifikasikan judul yang mirip dan tidak mirip. Ketiga, bagaimana cara menentukan nilai *threshold similarity* yang optimal agar diperoleh keseimbangan terbaik antara nilai *precision* dan *recall* dalam sistem deteksi kemiripan judul skripsi.

Sejalan dengan rumusan masalah tersebut, tujuan penelitian ini adalah untuk menganalisis performa algoritma Jaro–Winkler dan Levenshtein Distance dalam mendeteksi kemiripan judul skripsi. Selain itu, penelitian ini bertujuan untuk mengetahui dampak variasi nilai *threshold similarity* terhadap hasil klasifikasi sistem. Tujuan akhir dari penelitian ini adalah menentukan nilai ambang (*threshold*) yang optimal sehingga sistem mampu menghasilkan kinerja terbaik, yang diukur berdasarkan nilai F1-score atau Youden's Index sebagai indikator keseimbangan antara *precision* dan *recall*.

## **1.2 Pertanyaan Penelitian**

1. Bagaimana performa algoritma Jaro–Winkler dan Levenshtein dalam mendeteksi kemiripan judul skripsi?
2. Bagaimana pengaruh variasi nilai *threshold similarity* terhadap tingkat akurasi sistem penyaringan judul skripsi?
3. Bagaimana menentukan nilai *threshold similarity* optimal agar sistem dapat

mencapai keseimbangan terbaik antara precision dan recall?

### 1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang sudah dilampirkan, tujuan dari dilakukannya penelitian ini adalah:

1. Menganalisis performa algoritma Jaro–Winkler dan Levenshtein dalam mengukur

| Tahapan/<br>Stage                                  | Aktivitas yang<br>Dilakukan                                                                                                                                                                                                                                                  | Metode yang<br>Digunakan                                                                             | Output                                                                                                                                                                            |
|----------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1. <b>Persiapan Penelitian dan Studi Literatur</b> | <ol style="list-style-type: none"> <li>1. Studi pustaka mengenai algoritma string similarity (Jaro— Winkler, Levenshtein) dan konsep <i>threshold similarity</i>.</li> <li>2. Identifikasi permasalahan duplikasi judul skripsi dan kebutuhan sistem penyaringan.</li> </ol> | <ol style="list-style-type: none"> <li>1. Review Literatur</li> <li>2. Analisis Kebutuhan</li> </ol> | <ol style="list-style-type: none"> <li>1. Pemahaman mendalam tentang algoritma dan parameter threshold.</li> <li>2. Rumusan masalah dan batasan penelitian yang jelas.</li> </ol> |

| Tahapan/<br>Stage                                      | Aktivitas<br>yang<br>Dilakukan                                                                                                                                                                                               | Metode yang<br>Digunakan                           | Output                                                                             |
|--------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|------------------------------------------------------------------------------------|
| 2. <b>Pengumpulan<br/>dan<br/>Persiapan<br/>Data</b>   | <ol style="list-style-type: none"> <li>1. Menyusun dataset dummy berisi kumpulan judul skripsi fiktif dari berbagai bidang.</li> <li>2. Melakukan prapem rosesan teks (case folding, tokenisasi, normalisasi).</li> </ol>    | Simulasi dataset, text pre-processing              | Dataset siap uji dengan format standar dan label ground truth (mirip/tidak mirip). |
| 3. <b>Penghitungan<br/>Nilai<br/>Kemiripan</b>         | <ol style="list-style-type: none"> <li>1. Menghitung skor kemiripan antarjudul menggunakan algoritma Jaro—Winkler dan Levenshtein.</li> <li>2. Merekam hasil perhitungan untuk setiap pasangan judul.</li> </ol>             | Implementasi algoritma string similarity di Python | Nilai similarity score (0—1) dari kedua algoritma untuk setiap pasangan judul.     |
| 4. <b>Pengujian dan<br/>Perbandingan<br/>Threshold</b> | <ol style="list-style-type: none"> <li>1. Menerapkan beberapa nilai threshold (0.70—0.95) untuk menentukan batas kemiripan.</li> <li>2. Meng-klasifikasikan hasil (mirip/tidak mirip) berdasarkan nilai threshold</li> </ol> | Eksperimen komparatif                              | Tabel hasil klasifikasi dan performa sistem pada tiap nilai threshold.             |

| 5. <b>Evaluasi kinerja sistem</b>               | <ol style="list-style-type: none"> <li>1. Menghitung precision, recall, dan F1-score untuk tiap kombinasi algoritma dan threshold.</li> <li>2. Menentukan nilai ambang optimal berdasarkan F1-score tertinggi</li> </ol>     | Analisis kuantitatif (evaluasi metrik) | Nilai threshold optimal yang memberikan keseimbangan terbaik antara precision dan recall. |
|-------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------|-------------------------------------------------------------------------------------------|
| Tahapan/<br>Stage                               | Aktivitas yang Dilakukan                                                                                                                                                                                                     | Metode yang Digunakan                  | Output                                                                                    |
| 6. <b>Analisis Komparatif dan Uji Statistik</b> | Membandingkan performa Jaro—Winkler dan Levenshtein menggunakan Wilcoxon Signed-Rank                                                                                                                                         | Test. Uji statistik non-parametrik     | Hasil signifikansi perbedaan performa antar algoritma.                                    |
| 7. <b>Kesimpulan dan Rekomendasi</b>            | <ol style="list-style-type: none"> <li>1. Menarik kesimpulan dari hasil eksperimen dan analisis</li> <li>2. Memberikan saran untuk pengembangan sistem penyaringan</li> <li>3. judul skripsi berbasis similarity.</li> </ol> | Analisis deskriptif dan interpretatif  | Kesimpulan akhir dan rekomendasi pengembangan sistem                                      |