# Teknika 18 (2): first\_page – end\_page

JURNAL TEKNIKA

TEKNIKA

ISSN: 0854-3143

e-ISSN: 2622-3481

Journal homepage: http://jurnal.polsri.ac.id/index.php/teknika

Journal Email: teknika@polsri.ac.id

Kombinasi Algoritma TF-IDF dan Fuzzy Matching untuk Deteksi Kemiripan Judul Skripsi

Arif Nur Listanto1, Muhammad Fauzan Azima2, Fitria3, Chairani4

1, 2, 3, 4Program Studi Teknik Informatika, Institut Informatika Dan Bisnis Darmajaya, Lampung, Indonesia

\*Email Penulis Korespondensi: [arifnurlis.2011010035@mail.darmajaya.ac.id](mailto:arifnurlis.2011010035@mail.darmajaya.ac.id)

Abstrak

*Dalam era informasi digital, deteksi kemiripan teks menjadi penting untuk berbagai aplikasi seperti plagiarisme, pengelompokan dokumen, dan penyaringan informasi. Penelitian ini bertujuan untuk mengembangkan metode yang efektif dalam mendeteksi kemiripan teks dengan menggabungkan algoritma TF-IDF dan Fuzzy Matching. Alasan pemilihan topik ini didasarkan pada kebutuhan akan akurasi yang lebih tinggi dalam mengidentifikasi kemiripan teks yang seringkali tidak dapat diatasi dengan metode konvensional secara memadai. Metode penelitian ini melibatkan penggunaan TF-IDF untuk mengekstraksi fitur penting dari teks, yang kemudian dipadukan dengan Fuzzy Matching untuk mengatasi variasi dan ketidakpastian dalam teks yang dibandingkan. Hasil penelitian menunjukkan bahwa kombinasi kedua algoritma ini mampu meningkatkan akurasi deteksi kemiripan teks dibandingkan dengan penggunaan salah satu algoritma secara terpisah. Pengujian dilakukan pada berbagai dataset teks dengan variasi panjang dan kompleksitas, dan hasilnya menunjukkan peningkatan signifikan dalam tingkat keakuratan dan keandalan deteksi. Kesimpulan penelitian ini menegaskan pentingnya pendekatan gabungan TF-IDF dan Fuzzy Matching dalam aplikasi deteksi kemiripan teks, yang dapat memberikan kontribusi signifikan terhadap peningkatan kualitas dan efisiensi pengelolaan informasi digital.*

***Kata kunci****— Kemiripan Teks, TF-IDF, Fuzzy Matching, Deteksi Plagiarisme*

Abstract

*In the era of digital information, text similarity detection becomes important for various applications such as plagiarism, document clustering, and information filtering. This research aims to develop an effective method in detecting text similarity by combining TF-IDF and Fuzzy Matching algorithms. The reason for choosing this topic is based on the need for higher accuracy in identifying text similarity which often cannot be adequately addressed by conventional methods. The research method involves the use of TF-IDF to extract important features from the text, which is then combined with Fuzzy Matching to address variations and uncertainties in the compared texts. The results show that the combination of these two algorithms is able to improve the accuracy of text similarity detection compared to the use of either algorithm separately. Tests were conducted on various text datasets with variations in length and complexity, and the results showed significant improvements in detection accuracy and reliability. The conclusion of this research confirms the importance of the combined approach of TF-IDF and Fuzzy Matching in text similarity detection applications, which can significantly contribute to improving the quality and efficiency of digital information management.*

***Keywords—*** *Text Similarity, TF-IDF, Fuzzy Matching, plagiarism detection*

# 1. PENDAHULUAN

D

alam era digital yang semakin berkembang, pemrosesan teks telah menjadi bidang yang sangat penting, terutama dalam pendeteksian plagiarisme, pencarian informasi dan analisis sentiment. Dalam suatu instansi akademik, Penelitian atau biasa dikenal dengan istilah tugas akhir bagi mahasiswa merupakan sks wajib yang tersaji pada KRS setiap program studi atau jurusan di seluruh perguruan tinggi di dunia[1].

Salah satu syarat diterimanya sebuah judul skripsi adalah tingkat kesamaan judul skripsi yang diajukan dengan judul skripsi yang telah ada sebelumnya[2]. Setiap judul skripsi yang diajukan oleh mahasiswa akan di filter oleh ketua Kelompok Bidang Keilmuan (KBK). Di Darmajaya khususnya prodi Teknik Informatika terdapat 3 konsenterasi bidang keilmuan yaitu Artificial Intelligence, Mobile Technology, dan Multimedia.

Masalah yang sering terjadi yaitu judul yang diajukan oleh mahasiswa seringkali identik, sangat mirip dengan judul yang sudah pernah diajukan sebelumnya, ataupun mahasiswa hanya mengganti judul dengan lokasi penelitian yang berbeda tanpa adanya nilai keterbaharuan dari penelitian tersebut.

Penelitian oleh [3] membahas penggunaan algoritma Winnowing untuk mendeteksi plagiarisme dalam judul skripsi. Algoritma ini mencari kesamaan kata atau kalimat dengan membandingkan ‘fingerprint’ dari teks. Penelitian ini Menggunakan teknik hashing, algoritma mengubah n-gram teks menjadi nilai hash yang kemudian dibandingkan untuk menemukan kesamaan. Sistem yang diusulkan oleh penelitian ini memungkinkan pengecekan judul skripsi secara otomatis, membantu pengambilan keputusan apakah judul diterima atau ditolak.

Penelitian oleh [2] berfokus pada pembuatan sistem berbasis Android untuk mendeteksi persentase kemiripan judul skripsi menggunakan Algoritma Smith Waterman. Penelitian ini membangun sistem deteksi kemiripan judul skripsi untuk membantu mahasiswa mengetahui persentase kemiripan judul yang akan diajukan. Menggunakan Algoritma Smith Waterman untuk menganalisis data judul skripsi mahasiswa Teknik Informatika FIKOM UMI tahun 2018-2019. Penelitian ini berhasil mendeteksi kecocokan judul skripsi dengan presentase kemiripan tertinggi antara judul yang sama, serta menunjukkan persentase kemiripan yang lebih rendah untuk judul yang berbeda.

Penelitian oleh [4] berkaitan dengan pengembangan sistem deteksi tingkat kemiripan judul skripsi menggunakan algoritma Oliver, yang merupakan fungsi dalam bahasa pemrograman PHP untuk menghasilkan nilai kemiripan antara dua string1. Sistem ini dirancang untuk mengatasi masalah redundansi judul skripsi di universitas dengan membandingkan judul yang diajukan mahasiswa terhadap database judul skripsi yang sudah ada. Dengan menggunakan data uji dan data training, sistem mampu menolak judul dengan kemiripan di atas 60% dan menerima judul dengan kemiripan 60% atau kurang, membantu memastikan keunikan judul skripsi dan mengurangi risiko plagiarisme. Penelitian ini menunjukkan efektivitas algoritma Oliver dalam mendeteksi kemiripan judul skripsi dan memberikan solusi otomatis untuk proses validasi judul skripsi yang lebih cepat dan akurat.

# 2. METODE PENELITIAN

# A. Teori Algoritma *TF-IDF*

*TF-IDF* adalah sebuah metode yang merupakan integrasi antar *term frequency (TF)*, dan *inverse document frequency (IDF)*. Term Frequency dihitung menggunakan Persamaan dengan term frequency ke-i adalah frekuensi kemunculan term ke-i dalam dokumen ke-j. *Inverse Document Frequency (IDF)* adalah logaritma dari rasio jumlah seluruh dokumen dalam korpus dengan jumlah dokumen yang memiliki *term* yang dimaksud seperti yang dituliskan secara matematis pada Persamaan[5][6]. Algoitma *TF-IDF* dalam rumusnya dituliskan sebagai berikut:

Pada *TF (Term Frequency)* rumusnya adalah sebagai berikut:

*tf(t, d) =*  (1)

dimana:

t : *term*

d : *document*

Sedangkan *IDF (Inverse Document Frequency)* rumusnya adalah sebagai berikut:

*idf(t, D) =* (2)

dimana:

t : *term*

d : *document*

D : *total document*

N : *number of document*

Dan untuk rumus keseluruna dari *TF-IDF* sebagai berikut:

*tfidf(t, d, D) = tf(t, d) \* idf(t, D)* (3)

## B. Teori Algoritma Fuzzy Matching

*Fuzzy Matching* adalah teknik yang digunakan untuk mengidentifikasi tingkat kesamaan antara string, yang sangat berguna untuk penyelarasan ontologi di mana pencocokan yang tepat tidak selalu memungkinkan karena adanya variasi data. Makalah ini membahas beberapa algoritma untuk pencocokan *string* *fuzzy*, termasuk *Jaro-Winkler dan Levenshtein*[7][8]. Dalam penggunaan algoritma *Fuzzy Matching* terdapat beberapa langkah diantaranya:

1. *Preprocessing*:

Langkah ini dilakukan untuk mengubah huruf menjadi kecil, menghindari masalah dengan *case sensitive.*

2. Menghitung *Similarity Score*

Dalam langkah ini, merupakan metode untuk menghitung seberapa mirip dua string, beberapa metode yang umum digunakan diantarnya *Levenshtein Distance, Jaro-Winkler Distance* dan *Cosine Similarity.*

3. Membandingkan *String*

Langkah ini membandingkan *string* yang sudah dihitung *similarity score* pada lnagkah sebelumnya dan menemukan atau memilih nilai *similarity score* tertinggi dalam pencocokan *string.*

## C. Alur Penelitian

Dalam proses penelitian ini merupakan implementasi dari kombinasi algoritma *TF-IDF (Term Frequency-Inverse Document Frequency)* untuk mendeteksi kemiripan judul skripsi dan menggambarkan proses-proses secara jelas di dalam sistem.

Pada perancangan sistem ini juga membahas rancangan tampilan aplikasi yang akan dibangun. Alur *(Flowchart)* dari sistem pendeteksi kemiripan judul skripsi adalah sebagai berikut:

Input Judul

*Scan* Judul

Mulai

*Preprocessing Text (casefolding / lowercase)*

Judul Skripsi

*Preprocess Text (Tokenization)*

*Calculate TF-IDF Score Multiple By Fuzzy Matching Score*

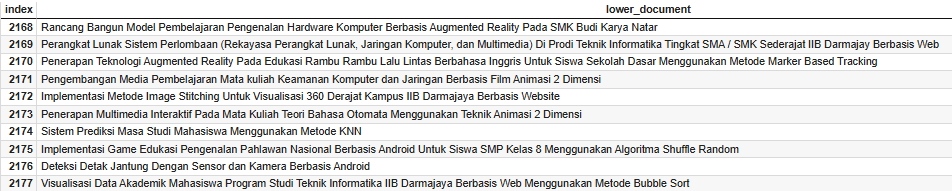
Analisis dan Evaluasi Hasil

Selesai

Gambar 1 Alur Penelitian *(Flowchart)* Sistem

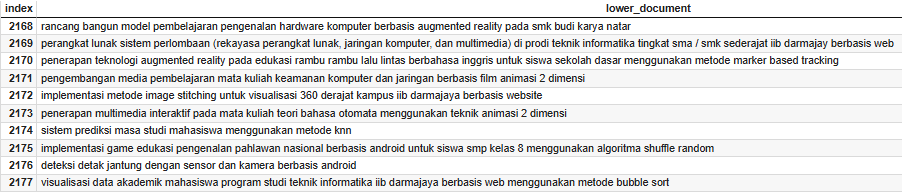
Berdasarkan alur penelitian pada gambar 1, terlihat bahwa pada langkah pertama pengguna diharuskan meng*input*kan judul yang akan dilakukan deteksi kemiripan dengan mengetikan secara manual pada *input* yang tersedia di dalam sistem.

Tabel 1 Sample Data Akhir Judul Skripsi

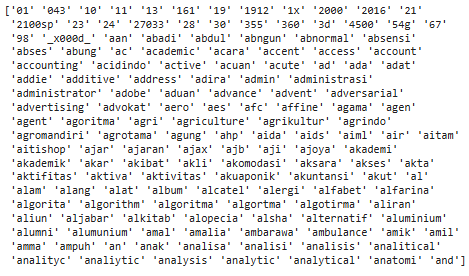


Tahap pertama yang dilakukan adalah memanggil semua data judul skripsi di dalam database ke dalam sistem, dapat terlihat pada table 1. Langkah selanjutnya yaitu mengubah semua huruf menjadi huruf kecil atau *casefolding (lowercase)* yang ditunjukan pada table 2.

Tabel 2 Hasil *CaseFolding (LowerCase)*



Setelah data dilakukan proses *casefolding (lowercase)* yang dapat terlihat pada tabel 2, langkah selanjutnya yaitu dilakukan proses *tokenization* yang dapat terlihat pada gambar 2.



Gambar 2 Data hasil proses *Tokenization* 150 Token

Langkah selanjutnya setelah proses *tokenization* adalah menghitung nilai *TF-IDF* dan dikalikan dengan nilai *Fuzzy Matching* terhadap judul yang di*input*kan oleh pengguna. Rumusnya dalam melakukan perhitungan kombinasi algoritma *TF-IDF* dengan *Fuzzy Matching* adalah sebagai berikut:

*F(x) =* (4)

*Score* atau nilai yang dihasilkan berupa bilangan bulat atau bilangan desimal. Sebagai contoh kata “kucing”, dengan kata “kulit” memiliki *similarity score* sekitar 0.55, apabila dikalikan dengan 100, maka akan mendapatkan nilai persentase sekitar 55%.

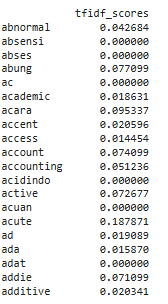
# 3. HASIL DAN PEMBAHASAN

Pembahasan terhadap hasil dan pengujian yang diperoleh dari peneilitian ini adalah sebagai berikut:

A. Proses Perhitungan *Similarity Score* Pada Input Terhadap Data

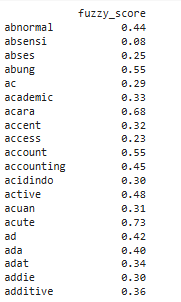
Dalam penelitian ini, peneliti menggunakan *dataset* Judul Skripsi Program Studi Teknik Informatika, IIB Darmajaya, dengan jumlah sekitar ±2148 judul skripsi pada *spreadsheet.* Berikut adalah hasil dari perhitungan nilai *TF-IDF* terhadap *dataset* dengan input “*perancangan sistem pakar untuk penyakit katarak*”

Tabel 3 Hasil *TF-IDF Score*

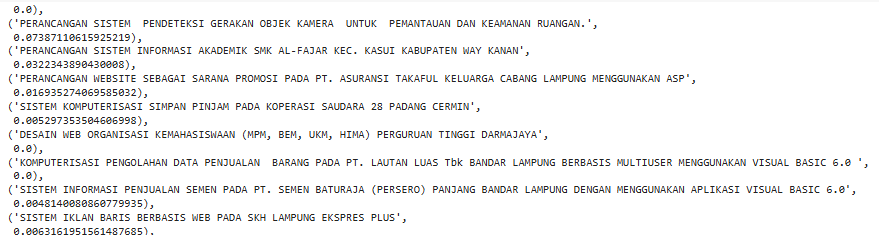


Dapat terlihat pada tabel 3, untuk input “*perancangan sistem pakar untuk penyakit katarak*” menghasilkan *score* seperti pada tabel 3 terhadap bebarapa *vector TF-IDF*. Langkah selanjutnya yaitu menghitung nilai *Fuzzy Matching* pada tabel 4terhadap input “*perancangan sistem pakar untuk penyakit katarak*”*.*

Tabel 4 Hasil Fuzzy Matching



Berdasarkan tabel 3 dan tabel 4, didapatkan *score* untuk masing-masing algoritma, sehingga langkah terakhir ialah menggabungkan kedua algoritma tersebut dengan cari dikalikan *matrix score* dari kedua algoritma tersebut, maka didapatkanlah hasil seperti berikut.



Gambar 3 Hasil dari *score* *similarty* terhadap *dataset*

Berdasarkan pada gambar 3, didapatkan nilai / *score* dengan input yang dimasukkan oleh user terhadap *dataset* Judul Skripsi Prodi Teknik Informatika IIB Darmajaya.

B. Merancang Aplikasi *Android*

Dalam penelitian ini, peneliti merancang hasil dari kombinasi algoritma *TF-IDF* dan *Fuzzy Matching* dalam bentuk aplikasi *Android*, agar dapat digunakan untuk melakukan proses pengecekan judul skripsi.

# 4. KESIMPULAN

Kesimpulan harus mengindikasi secara jelas hasil-hasil yang diperoleh, kelebihan dan kekurangannya, serta kemungkinan pengembangan selanjutnya.

Kesimpulan dapat berupa paragraf, namun sebaiknya berbentuk point-point dengan menggunakan numbering atau bullet.

# 5. SARAN

Saran-saran untuk untuk penelitian lebih lanjut untuk menutup kekurangan penelitian. Tidak memuat saran-saran diluar untuk penelitian lanjut.

# UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Tim Redaksi Jurnal Teknika Politeknik Negeri Sriwijaya yang telah memberi memberi kesempatan, sehingga artikel ilmiah ini dapat diterbitkan.

# DAFTAR PUSTAKA

[1] Mauliddin, “Analisis Kesalahan Penggunaan Uji Statistik pada Skripsi Mahasiswa,” *J. Mat. dan Pembelajaran*, 2017.

[2] U. N. Hasanah, R. Satra, and F. Umar, “Deteksi Kemiripan Judul Skripsi Menggunakan Algoritma Smith Waterman,” *Bul. Sist. Inf. dan Teknol. Islam*, 2020, doi: 10.33096/busiti.v1i1.676.

[3] N. Alamsyah, “DETEKSI PLAGIARISME TINGKAT KEMIRIPAN JUDUL SKRIPSI DENGAN ALGORITMA WINNOWING,” *Technol. J. Ilm.*, 2017, doi: 10.31602/tji.v8i4.1119.

[4] N. Fitrianingsih, M. Asfi, D. Prasetyo, R. P. Kusuma, and M. A. Sulhan, “Deteksi Tingkat Kemiripan Judul Menggunakan Algoritma Oliver Pada Sistem Informasi Pengajuan Skripsi,” *J. MEDIA Inform. BUDIDARMA*, 2022, doi: 10.30865/mib.v6i4.4409.

[5] Y. Gu, Y. Wang, J. Huan, Y. Sun, and S. Xu, “An improved TFIDF algorithm based on dual parallel adaptive computing model,” *Int. J. Embed. Syst.*, 2020, doi: 10.1504/IJES.2020.108278.

[6] N. K. Widyasanti, I. K. G. Darma Putra, and N. K. Dwi Rusjayanthi, “Seleksi Fitur Bobot Kata dengan Metode TFIDF untuk Ringkasan Bahasa Indonesia,” *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, 2018, doi: 10.24843/jim.2018.v06.i02.p06.

[7] T. Efriyanto and M. Hayaty, “JARO WINKLER ALGORITHM FOR MEASURING SIMILARITY ONLINE NEWS,” *J. Tek. Inform.*, 2022.

[8] B. Berger, M. S. Waterman, and Y. W. Yu, “Levenshtein Distance, Sequence Comparison and Biological Database Search,” *IEEE Trans. Inf. Theory*, 2021, doi: 10.1109/TIT.2020.2996543.