CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

Analisis Topik dan Karakteristik Kelayakan Pengajuan Judul Skripsi Mahasiswa Program Studi Sistem Informasi Menggunakan TF-IDF dan K-Means Clustering

Singgih Yulizar Ma'ruf[1*], David Naista[2]
[1,3]Sistem Informasi, Universitas Islam Negeri Raden Intan Lampung, Indonesia
[1*]singgihyulizarm@radenintan.ac.id, [2]davidnaista@radenintan.ac.id

Abstrak: Pengajuan judul skripsi merupakan tahapan awal yang menentukan arah penelitian mahasiswa. Banyaknya variasi judul yang diajukan setiap tahun menyebabkan proses identifikasi tren penelitian dan evaluasi karakteristik kelayakan judul menjadi semakin kompleks apabila dilakukan secara manual. Penelitian ini bertujuan untuk menganalisis topik penelitian serta karakteristik kelayakan pengajuan judul skripsi mahasiswa Program Studi Sistem Informasi menggunakan metode TF-IDF dan K-Means Clustering. Data yang digunakan berasal dari 757 judul skripsi unik yang tersimpan pada sistem pengajuan judul skripsi. Tahapan penelitian meliputi preprocessing teks, pembobotan kata menggunakan TF-IDF, pembentukan cluster menggunakan algoritma K-Means, serta analisis karakteristik kelayakan berdasarkan similarity score, skor kelayakan, dan status pengajuan. Penentuan jumlah cluster dilakukan menggunakan kombinasi Elbow Method dan Silhouette Score sehingga diperoleh enam cluster yang dapat diinterpretasikan secara tematik. Hasil penelitian menunjukkan bahwa topik penelitian mahasiswa terkonsentrasi pada enam kelompok utama, yaitu sistem informasi administrasi dan manajemen, aplikasi mobile, sistem informasi desa, sistem pendukung keputusan, monitoring berbasis dashboard dan Internet of Things (IoT), serta sistem informasi pariwisata. Analisis karakteristik kelayakan menunjukkan bahwa cluster monitoring dan IoT memiliki rata-rata skor kelayakan tertinggi sebesar 92,36 dengan persentase status layak mencapai 80%, sedangkan cluster pariwisata memiliki rata-rata skor kelayakan terendah sebesar 79,47 dan persentase revisi tertinggi sebesar 70,91%. Hasil pengujian Kruskal-Wallis menghasilkan p-value sebesar $1{,}51\times10^{-9}$ yang menunjukkan adanya perbedaan signifikan skor kelayakan antar cluster. Temuan ini menunjukkan bahwa topik penelitian memiliki hubungan dengan tingkat kelayakan pengajuan judul skripsi dan dapat dimanfaatkan sebagai bahan evaluasi serta pengambilan keputusan akademik berbasis data.
Kata Kunci: Text Mining; TF-IDF; K-Means Clustering; Judul Skripsi; Kelayakan Judul

Abstract: The increasing number of thesis title submissions in the Information Systems Study Program has created challenges in identifying research trends and evaluating the feasibility characteristics of proposed topics. Most previous studies on document clustering focus primarily on topic grouping without

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

considering the quality or feasibility attributes associated with each document. Therefore, this study aims to analyze research topic patterns and feasibility characteristics of undergraduate thesis title submissions using the combination of Term Frequency–Inverse Document Frequency (TF-IDF) and K-Means Clustering. The dataset consisted of 757 unique thesis title submissions obtained from the thesis title proposal system of the Information Systems Study Program. The research process included text preprocessing, TF-IDF feature extraction, clustering using K-Means, cluster evaluation using Silhouette Score and Elbow Method, visualization using t-distributed Stochastic Neighbor Embedding (t-SNE), and feasibility analysis based on similarity score, feasibility score, and submission status. Statistical testing was further conducted using the Kruskal–Wallis test to determine whether significant differences existed among clusters. The clustering process produced six interpretable topic groups representing administrative information systems, mobile applications, village information systems, decision support systems, monitoring and Internet of Things (IoT), and tourism information systems. The highest average feasibility score was found in the monitoring and IoT cluster (92.36), while the tourism-related cluster obtained the lowest average score (79.47) and the highest revision rate. The Kruskal–Wallis test produced a p-value of 1.51×10^{-9}, indicating significant differences in feasibility scores among clusters. The novelty of this study lies in integrating topic clustering with feasibility characteristic analysis of thesis title submissions. The results provide valuable insights for academic program management, research topic planning, and data-driven decision-making in thesis title evaluation.
Keywords: Text Mining; TF-IDF; K-Means Clustering; Thesis Title Analysis; Feasibility Assessment

1. PENDAHULUAN

Skripsi merupakan salah satu bentuk karya ilmiah yang wajib diselesaikan oleh mahasiswa sebagai syarat memperoleh gelar sarjana. Dalam proses penyusunan skripsi, tahap pengajuan judul menjadi langkah awal yang sangat penting karena menentukan arah penelitian yang akan dilakukan. Setiap tahun Program Studi Sistem Informasi Fakultas Sains dan Teknologi UIN Raden Intan Lampung menerima berbagai usulan judul skripsi dengan tema yang beragam, mulai dari sistem informasi, data mining, kecerdasan buatan, sistem pendukung keputusan, aplikasi mobile, hingga pengembangan website. Banyaknya variasi judul yang diajukan menyebabkan proses identifikasi tren penelitian menjadi semakin kompleks apabila dilakukan secara manual.

Di sisi lain, ketersediaan data historis pengajuan judul skripsi sebenarnya dapat dimanfaatkan untuk memperoleh informasi mengenai kecenderungan topik penelitian mahasiswa. Informasi tersebut dapat digunakan oleh program studi untuk mengevaluasi pemerataan bidang penelitian, mengidentifikasi topik yang terlalu dominan, serta memberikan rekomendasi kepada mahasiswa mengenai peluang penelitian yang masih terbuka. Namun, jumlah data yang besar dan berbentuk teks tidak terstruktur menyebabkan proses analisis membutuhkan pendekatan yang mampu mengolah dokumen secara otomatis dan sistematis.

Perkembangan bidang text mining memberikan berbagai metode untuk mengekstraksi informasi dari kumpulan dokumen teks dalam jumlah besar. Text mining memungkinkan proses pengolahan data tekstual melalui tahapan preprocessing, transformasi data, ekstraksi fitur, hingga analisis pola yang terkandung dalam dokumen. Salah satu teknik yang banyak digunakan adalah Term Frequency-Inverse Document Frequency (TF-IDF), yaitu metode pembobotan kata yang mampu merepresentasikan dokumen ke dalam bentuk numerik berdasarkan tingkat kepentingan suatu kata dalam keseluruhan koleksi

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

dokumen [1][2]. Representasi TF-IDF telah banyak diterapkan pada berbagai penelitian pengelompokan dokumen karena mampu menghasilkan fitur yang relevan dan efisien untuk proses clustering [3].

Setelah dokumen direpresentasikan dalam bentuk vektor numerik, proses pengelompokan dapat dilakukan menggunakan algoritma clustering. Salah satu algoritma clustering yang paling populer adalah K-Means karena memiliki tingkat kompleksitas yang relatif rendah, mudah diimplementasikan, dan mampu menangani data dalam jumlah besar[4][5]. Algoritma K-Means bekerja dengan mengelompokkan dokumen yang memiliki karakteristik serupa ke dalam cluster yang sama sehingga pola atau tema tertentu dapat diidentifikasi secara lebih mudah. Dalam konteks analisis dokumen, kombinasi TF-IDF dan K-Means telah banyak digunakan untuk mengelompokkan berita, artikel ilmiah, dokumen akademik, maupun data media sosial [6][7].

Beberapa penelitian terdahulu menunjukkan bahwa kombinasi pembobotan TF-IDF dan algoritma K-Means efektif digunakan untuk mengelompokkan dokumen teks berdasarkan kemiripan kata dan topik yang terkandung di dalamnya. Penelitian yang dilakukan oleh Siahaan dkk. menunjukkan bahwa TF-IDF dan K-Means mampu mengidentifikasi topik dominan pada kumpulan publikasi ilmu komputer sehingga dapat digunakan untuk memetakan tren penelitian secara otomatis[1]. Penelitian lain menerapkan algoritma K-Means pada koleksi perpustakaan dan menunjukkan bahwa proses pengelompokan mampu membantu pengelolaan serta aksesibilitas informasi berdasarkan karakteristik koleksi yang dimiliki[8]. Selain itu, penelitian oleh Zendrato dkk. berhasil memanfaatkan metode clustering untuk mengelompokkan dokumen tugas akhir mahasiswa berdasarkan tema penelitian sehingga memberikan gambaran distribusi bidang kajian pada program studi yang diteliti[6].

Meskipun demikian, sebagian besar penelitian sebelumnya hanya berfokus pada proses pengelompokan topik dokumen menggunakan pendekatan text mining dan clustering. Hasil penelitian umumnya berhenti pada identifikasi kelompok topik atau tren penelitian tanpa melakukan analisis terhadap karakteristik kualitas dokumen yang berada pada setiap kelompok yang terbentuk. Pada konteks pengajuan judul skripsi, informasi mengenai kualitas usulan judul merupakan aspek penting yang dapat digunakan untuk mendukung proses evaluasi akademik dan pengambilan keputusan oleh program studi.

Berbeda dengan penelitian terdahulu, penelitian ini tidak hanya melakukan pengelompokan judul skripsi berdasarkan kemiripan topik menggunakan kombinasi TF-IDF dan K-Means Clustering, tetapi juga mengintegrasikan atribut penilaian yang tersedia pada sistem pengajuan judul skripsi, yaitu similarity score, skor kelayakan, dan status pengajuan. Melalui pendekatan tersebut, setiap cluster tidak hanya merepresentasikan kelompok topik penelitian, tetapi juga menunjukkan karakteristik tingkat kelayakan dari usulan judul yang berada pada cluster tersebut. Selain itu, penelitian ini melakukan pengujian statistik menggunakan Kruskal–Wallis untuk mengevaluasi perbedaan skor kelayakan antar cluster yang terbentuk.

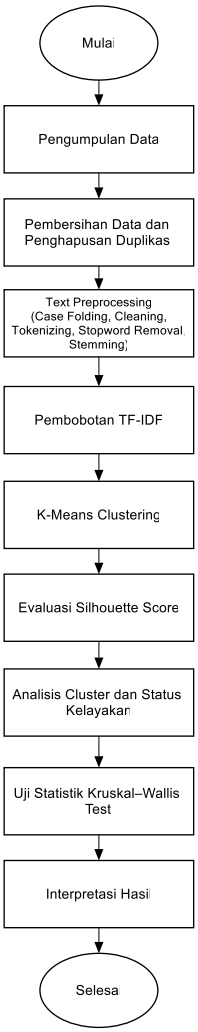
Novelty penelitian ini terletak pada integrasi analisis topik dan analisis kelayakan pengajuan judul skripsi dalam satu kerangka penelitian berbasis text mining. Berbeda dengan penelitian clustering dokumen sebelumnya yang umumnya hanya menghasilkan kelompok topik, penelitian ini memanfaatkan atribut similarity score, skor kelayakan, dan status pengajuan untuk mengidentifikasi karakteristik kualitas pada setiap cluster yang terbentuk. Selain itu, penelitian ini melakukan pengujian statistik menggunakan Kruskal–Wallis untuk menguji perbedaan tingkat kelayakan antar kelompok topik. Dengan demikian, penelitian tidak hanya menghasilkan pemetaan topik penelitian mahasiswa, tetapi juga memberikan informasi mengenai hubungan antara topik penelitian dan kecenderungan kelayakan usulan judul skripsi.

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan text mining dan clustering untuk menganalisis pola topik serta karakteristik kelayakan pengajuan judul skripsi mahasiswa Program Studi Sistem Informasi. Metode yang digunakan terdiri dari beberapa tahapan, yaitu pengumpulan data, preprocessing teks, pembobotan TF-IDF, proses clustering menggunakan algoritma K-Means, evaluasi cluster menggunakan Silhouette Coefficient, serta analisis karakteristik kelayakan berdasarkan hasil pengelompokan. Alur penelitian yang digunakan dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

Data penelitian diperoleh dari sistem pengajuan judul skripsi yang digunakan oleh Program Studi Sistem Informasi Fakultas Sains dan Teknologi UIN Raden Intan Lampung. Dataset terdiri atas 950 data pengajuan judul skripsi yang diajukan mahasiswa selama periode pengamatan. Setelah proses pembersihan duplikasi, diperoleh 757 judul unik yang digunakan sebagai objek analisis. Selain informasi judul, dataset juga memuat atribut status pengajuan (layak, revisi, dan ditolak), similarity score, serta skor kelayakan yang diberikan oleh sistem.

Tahap pertama adalah preprocessing teks untuk mengubah dokumen menjadi bentuk yang siap dianalisis. Tahapan preprocessing meliputi case folding, tokenisasi, penghapusan stopword, dan stemming menggunakan algoritma Bahasa Indonesia. Tahapan ini

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

bertujuan mengurangi variasi kata yang memiliki makna sama sehingga representasi dokumen menjadi lebih konsisten. Proses preprocessing merupakan tahapan penting dalam text mining karena dapat meningkatkan kualitas representasi fitur dokumen dan hasil clustering yang dihasilkan [9].

Setelah preprocessing, setiap dokumen direpresentasikan menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF digunakan untuk memberikan bobot pada kata berdasarkan tingkat kemunculannya dalam dokumen dan keseluruhan koleksi dokumen. Kata yang sering muncul pada suatu dokumen tetapi jarang muncul pada dokumen lain akan memperoleh bobot yang lebih tinggi sehingga mampu merepresentasikan topik dokumen secara lebih baik. Pendekatan TF-IDF telah banyak digunakan dalam analisis dokumen dan clustering teks karena mampu menghasilkan representasi fitur yang efektif pada data berdimensi tinggi [10].

Representasi TF-IDF kemudian digunakan sebagai input algoritma K-Means Clustering. K-Means merupakan metode unsupervised learning yang mengelompokkan dokumen berdasarkan tingkat kemiripan fitur sehingga dokumen dengan karakteristik serupa berada pada cluster yang sama. Algoritma ini dipilih karena memiliki performa yang baik pada data teks berukuran besar, sederhana dalam implementasi, serta banyak digunakan dalam penelitian text mining dan document clustering [11].

Tahapan clustering dilakukan dengan menguji beberapa jumlah cluster (k) dan mengevaluasi kualitas hasil pengelompokan menggunakan Silhouette Coefficient. Nilai Silhouette mengukur tingkat kedekatan anggota dalam suatu cluster dibandingkan dengan cluster lainnya. Semakin tinggi nilai Silhouette, semakin baik kualitas pemisahan antar cluster yang dihasilkan. Pada penelitian ini diperoleh nilai Silhouette tertinggi sebesar 0,0314 pada k = 6 sehingga jumlah cluster yang digunakan dalam analisis akhir adalah enam cluster. Penggunaan Silhouette Coefficient sebagai metode evaluasi internal clustering telah banyak digunakan dalam penelitian clustering karena mampu menggambarkan tingkat kohesi dan separasi cluster secara simultan [12].

Tahap terakhir adalah analisis karakteristik setiap cluster. Analisis dilakukan dengan mengidentifikasi kata kunci dominan pada masing-masing cluster menggunakan nilai TF-IDF tertinggi serta menghubungkannya dengan informasi status pengajuan, similarity score, dan skor kelayakan. Selanjutnya dilakukan pengujian statistik menggunakan Kruskal–Wallis Test untuk mengetahui apakah terdapat perbedaan signifikan skor kelayakan antar cluster yang terbentuk. Uji Kruskal–Wallis merupakan metode nonparametrik yang digunakan untuk membandingkan tiga kelompok atau lebih yang bersifat independen, terutama ketika data tidak memenuhi asumsi normalitas sebagaimana disyaratkan pada analisis varians (ANOVA) [13]. Pengujian ini dilakukan terhadap skor kelayakan pada setiap cluster hasil K-Means untuk mengidentifikasi apakah karakteristik topik yang berbeda memiliki tingkat kelayakan pengajuan judul yang berbeda secara statistik. Hasil analisis ini digunakan untuk mengidentifikasi kecenderungan topik penelitian mahasiswa serta hubungan antara topik penelitian dan tingkat kelayakan pengajuan judul skripsi.

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Dataset

Penelitian ini menggunakan data pengajuan judul skripsi mahasiswa Program Studi Sistem Informasi yang diperoleh dari sistem pengajuan judul skripsi. Dataset awal terdiri dari 950 data pengajuan. Setelah dilakukan proses pembersihan data dan penghapusan duplikasi judul, diperoleh 757 judul unik yang digunakan dalam proses analisis.

Tabel 1. Ringkasan Dataset Penelitian

Keterangan	Jumlah
Total data pengajuan	950

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

Judul unik	757
Status Layak	462
Status Revisi	260
Status Ditolak	35

Setiap data terdiri atas judul skripsi, similarity score, status pengajuan, dan skor kelayakan. Status pengajuan terdiri dari tiga kategori yaitu layak, revisi, dan ditolak. Informasi tersebut digunakan untuk menganalisis hubungan antara topik penelitian mahasiswa dengan tingkat kelayakan pengajuan judul skripsi.

3.2 Mekanisme Penilaian Similarity Score dan Skor Kelayakan

Dataset penelitian diperoleh dari sistem pengajuan judul skripsi Program Studi Sistem Informasi Fakultas Sains dan Teknologi UIN Raden Intan Lampung yang menyimpan informasi judul usulan mahasiswa beserta hasil evaluasi otomatis. Selain data judul, sistem juga menghasilkan atribut similarity score, skor kelayakan, dan status pengajuan yang digunakan dalam proses analisis penelitian ini.

3.2.1 Similarity Score

Similarity score digunakan untuk mengukur tingkat kemiripan antara judul yang diajukan mahasiswa dengan koleksi judul skripsi yang telah tersimpan sebelumnya pada basis data program studi. Pengukuran kemiripan dilakukan menggunakan kombinasi tiga metode, yaitu:

- Levenshtein Distance, untuk mengukur jumlah perubahan karakter yang diperlukan untuk mengubah satu judul menjadi judul lainnya.
- Similar Text, untuk menghitung tingkat kesamaan teks berdasarkan kecocokan karakter dan kata antar dokumen.
- Cosine Similarity, untuk mengukur kedekatan vektor representasi teks berdasarkan sudut antar dokumen.

Nilai akhir similarity score dinyatakan dalam rentang 0–100, di mana nilai yang lebih tinggi menunjukkan tingkat kemiripan yang lebih besar dengan judul yang telah ada sebelumnya.

3.2.2 Skor Kelayakan

Skor kelayakan digunakan untuk menilai kualitas dan potensi kelayakan usulan judul skripsi. Penilaian dilakukan berdasarkan tujuh indikator utama yang telah ditetapkan oleh program studi, yaitu:

- Masalah Nyata
- Analisis atau Algoritma
- Evaluasi
- Kompleksitas
- Teknologi Modern
- Metodologi
- Kebaruan

Masing-masing indikator dinilai pada rentang skor tertentu dan kemudian diakumulasikan menjadi nilai akhir dalam skala 0–100. Semakin tinggi nilai yang diperoleh, semakin besar tingkat kelayakan judul untuk dilanjutkan ke tahap penelitian.

Berdasarkan nilai similarity score dan skor kelayakan yang diperoleh, sistem menghasilkan status pengajuan yang terdiri atas tiga kategori, yaitu layak, revisi, dan ditolak. Pada penelitian ini, ketiga atribut tersebut digunakan untuk menganalisis karakteristik kelayakan pada setiap cluster yang terbentuk.

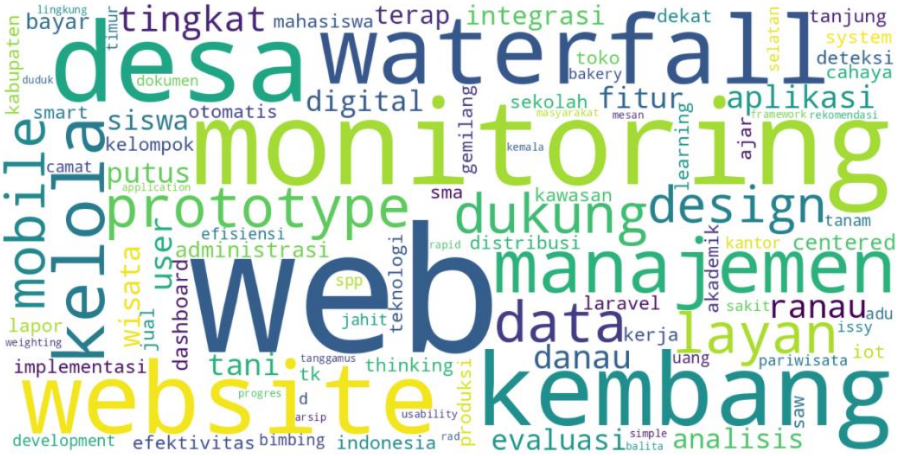
*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

3.3 Hasil Preprocessing dan Pembobotan TF-IDF

Tahap preprocessing dilakukan melalui proses case folding, tokenisasi, penghapusan stopword, dan stemming. Proses ini bertujuan untuk menghilangkan kata-kata yang tidak memiliki kontribusi signifikan terhadap pembentukan topik dokumen serta menyeragamkan bentuk kata.

Setelah preprocessing, dokumen direpresentasikan menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Pembobotan TF-IDF digunakan untuk mengukur tingkat kepentingan suatu kata terhadap dokumen dan keseluruhan koleksi dokumen. Hasil pembobotan menghasilkan sejumlah kata dominan yang sering muncul pada pengajuan judul skripsi mahasiswa.

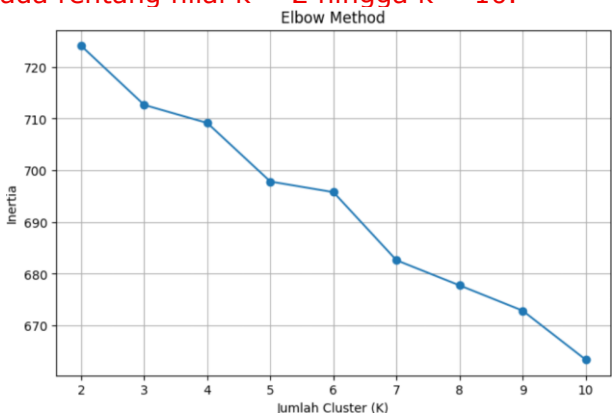


Gambar 2. Word Cloud Kata Dominan pada Dataset Judul Skripsi Mahasiswa

Beberapa kata yang memiliki frekuensi tinggi setelah preprocessing antara lain monitoring, desa, pengembangan, manajemen, pengelolaan, mobile, data, layanan, analisis, dan dashboard. Kata-kata tersebut menunjukkan kecenderungan topik penelitian mahasiswa yang berfokus pada pengembangan sistem informasi, digitalisasi layanan, monitoring, dan pengelolaan data.

3.3 Hasil Clustering Menggunakan K-Means

Proses clustering dilakukan menggunakan algoritma K-Means terhadap matriks TF-IDF yang diperoleh dari hasil preprocessing judul skripsi. Penentuan jumlah cluster dilakukan melalui kombinasi evaluasi Elbow Method, Silhouette Score, dan interpretasi topik cluster. Pengujian dilakukan pada rentang nilai k = 2 hingga k = 10.



Gambar 3. Elbow Method

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

Berdasarkan Gambar 4, nilai inertia mengalami penurunan yang cukup signifikan hingga rentang k=5 sampai k=6, kemudian cenderung menurun secara lebih landai pada jumlah cluster berikutnya. Kondisi tersebut menunjukkan bahwa jumlah cluster yang representatif berada pada rentang 5–7 cluster sehingga nilai k=6 dipilih sebagai jumlah cluster yang digunakan dalam penelitian ini.

Tabel 2. Silhouette Score

K	Silhouette Score
2	0.018
3	0.0245
4	0.0249
5	0.0325
6	0.0314
7	0.0429
8	0.0465
9	0.0496
10	0.0569

Selanjutnya dilakukan pengujian menggunakan Silhouette Coefficient pada rentang k=2 hingga k=10. Hasil pengujian menunjukkan bahwa nilai Silhouette Score pada k=6 sebesar 0,0314. Meskipun nilai tersebut bukan nilai tertinggi yang diperoleh, peningkatan nilai silhouette pada jumlah cluster yang lebih besar relatif kecil. Sebagai contoh, nilai silhouette tertinggi diperoleh pada k=10 sebesar 0,0569, namun menghasilkan jumlah cluster yang lebih banyak dan cenderung memecah topik menjadi kelompok yang lebih kecil sehingga mengurangi kemudahan interpretasi hasil clustering.

Dengan mempertimbangkan hasil Elbow Method, interpretabilitas topik yang dihasilkan, serta tujuan penelitian untuk memetakan kelompok tema penelitian mahasiswa, maka jumlah cluster sebanyak enam dipilih sebagai konfigurasi yang paling sesuai. Pemilihan jumlah cluster pada penelitian ini tidak hanya didasarkan pada nilai evaluasi statistik, tetapi juga mempertimbangkan makna substantif dari topik yang terbentuk. Tabel 3 menyajikan distribusi jumlah anggota pada masing-masing cluster yang terbentuk.

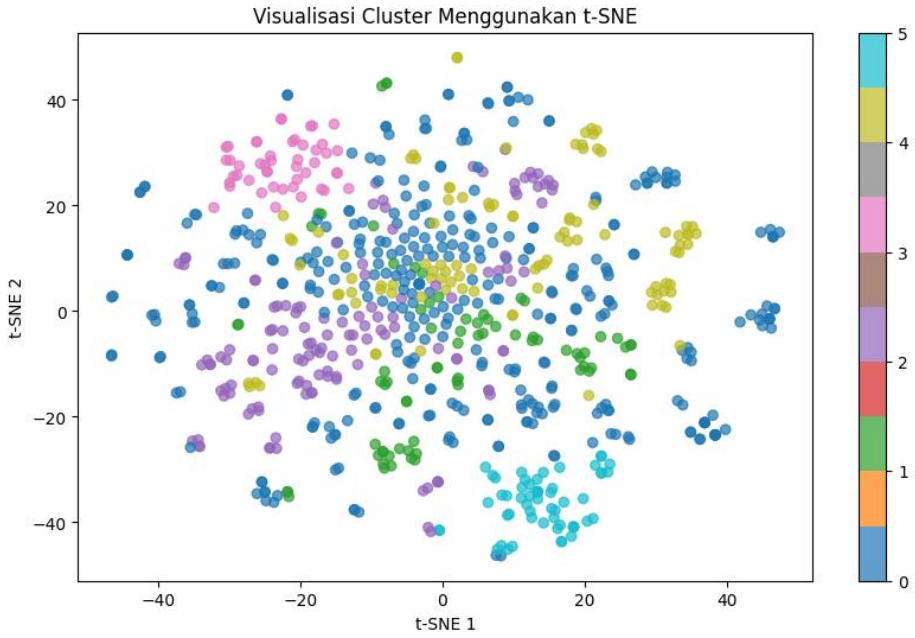
Tabel 3. Distribusi Anggota Cluster

Cluster	Jumlah Data
Cluster 0	353
Cluster 1	74
Cluster 2	128
Cluster 3	47
Cluster 4	100
Cluster 5	55
Total	757

Selain menggunakan Silhouette Score, penelitian ini juga melakukan evaluasi visual menggunakan teknik t-Distributed Stochastic Neighbor Embedding (t-SNE). Visualisasi t-SNE digunakan untuk memproyeksikan representasi TF-IDF berdimensi tinggi ke dalam ruang dua dimensi sehingga pola penyebaran cluster dapat diamati secara visual. Hasil visualisasi menunjukkan bahwa sebagian cluster membentuk area yang relatif terpisah meskipun masih terdapat beberapa area yang saling beririsan. Kondisi tersebut mengindikasikan bahwa struktur topik pada data pengajuan judul skripsi bersifat heterogen dan memiliki keterkaitan antar bidang penelitian.

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283



Gambar 4. Visualisasi Cluster Menggunakan t-SNE

Berdasarkan Gambar 4 terlihat bahwa cluster yang terbentuk tidak sepenuhnya terpisah secara tegas. Namun demikian, beberapa kelompok data menunjukkan kecenderungan membentuk area yang berdekatan sesuai karakteristik topiknya. Hasil ini mendukung temuan bahwa meskipun nilai Silhouette Score relatif rendah, cluster yang dihasilkan masih dapat digunakan untuk mengidentifikasi pola topik penelitian mahasiswa.

3.4 Karakteristik Topik Setiap Cluster

Karakteristik setiap cluster dianalisis berdasarkan kata-kata dengan bobot TF-IDF tertinggi. Hasil analisis menunjukkan bahwa setiap cluster memiliki kecenderungan topik yang berbeda.

Tabel 3. Ringkasan Tema Cluster

Cluster	Tema Dominan
0	Sistem Informasi Administrasi dan Manajemen
1	Mobile Application dan Klasifikasi Cerdas
2	Sistem Informasi Desa dan Pelayanan Publik
3	Sistem Pendukung Keputusan
4	Monitoring Dashboard dan IoT
5	Sistem Informasi Pariwisata

Cluster 0 didominasi oleh kata manajemen, pembayaran, spp, analisis, dan pengembangan yang menunjukkan fokus pada sistem administrasi dan manajemen pendidikan.
Cluster 1 didominasi oleh kata mobile, layanan, konseling, penyakit, dan deteksi yang menunjukkan fokus pada aplikasi mobile dan sistem cerdas berbasis klasifikasi.
Cluster 2 memiliki kata dominan desa, pengelolaan, pelayanan, distribusi, dan pendataan yang mengindikasikan topik sistem informasi desa dan pengelolaan data masyarakat.
Cluster 3 didominasi oleh kata pendukung keputusan, saw, pemilihan, dan topsis yang menunjukkan kelompok penelitian Sistem Pendukung Keputusan (SPK).

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

Cluster 4 berfokus pada monitoring, dashboard, IoT, laboratorium, dan pelabuhan yang menggambarkan penelitian monitoring berbasis dashboard dan Internet of Things.
Cluster 5 didominasi oleh kata danau ranau, wisata, pariwisata, user centered design, dan pengembangan wisata yang menunjukkan kelompok penelitian sistem informasi pariwisata.

3.5 Analisis Kelayakan Judul Berdasarkan Cluster

Tabel 4. Rata-rata Similarity Score dan Skor Kelayakan

Cluster	Mean Similarity	Mean Kelayakan
0	44.45	84.24
1	41.95	90.24
2	44.94	90.41
3	49.39	87.53
4	41.66	92.36
5	52.85	79.47

Hasil analisis menunjukkan bahwa Cluster 4 memiliki rata-rata skor kelayakan tertinggi yaitu 92,36. Cluster ini didominasi topik monitoring, dashboard, dan Internet of Things yang menunjukkan tingkat penerimaan yang tinggi dalam proses pengajuan judul skripsi.

Sebaliknya, Cluster 5 memiliki rata-rata skor kelayakan terendah yaitu 79,47 meskipun memiliki similarity score tertinggi sebesar 52,85. Temuan ini menunjukkan bahwa tingginya kemiripan judul dengan penelitian sebelumnya tidak selalu berbanding lurus dengan tingkat kelayakan yang diberikan.

Hasil tersebut mengindikasikan bahwa aspek kebaruan dan relevansi topik penelitian turut memengaruhi keputusan kelayakan pengajuan judul.

Tabel 5. Persentase Status Kelayakan per Cluster

Cluster	Layak (%)	Revisi (%)	Ditolak (%)
0	53.26	39.09	7.65
1	85.14	13.51	1.35
2	70.31	28.91	0.78
3	63.83	34.04	2.13
4	80	20	0
5	20	70.91	9.09

3.6 Pengujian Statistik Kruskal-Wallis

Untuk mengetahui apakah terdapat perbedaan skor kelayakan antar cluster, dilakukan pengujian menggunakan Kruskal-Wallis Test.

Tabel 6. Hasil Uji Kruskal-Wallis

Parameter	Nilai
Statistik Uji	49.813
P-value	1.51×10^{-9}
Keputusan	H_0 Ditolak

Hasil pengujian menunjukkan nilai statistik sebesar 49,813 dengan p-value sebesar $1{,}51 \times 10^{-9}$. Nilai p-value yang jauh lebih kecil dibandingkan tingkat signifikansi 0,05

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

menunjukkan bahwa terdapat perbedaan skor kelayakan yang signifikan antar cluster yang terbentuk.

Hasil ini menunjukkan bahwa karakteristik topik penelitian memiliki hubungan yang signifikan terhadap tingkat kelayakan pengajuan judul skripsi mahasiswa.

3.7 Pembahasan

Hasil clustering menunjukkan bahwa pengajuan judul skripsi mahasiswa Program Studi Sistem Informasi Fakultas Sains dan Teknologi UIN Raden Intan Lampung memiliki kecenderungan terkonsentrasi pada beberapa tema utama, yaitu sistem informasi administrasi dan manajemen, aplikasi mobile, sistem informasi desa, sistem pendukung keputusan, monitoring berbasis dashboard dan Internet of Things (IoT), serta sistem informasi pariwisata.

Cluster dengan jumlah anggota terbesar adalah Cluster 0 yang berisi 353 dokumen. Dominasi cluster ini menunjukkan bahwa penelitian terkait sistem informasi administrasi, manajemen data, pembayaran, dan pengelolaan organisasi masih menjadi pilihan utama mahasiswa. Kondisi ini sejalan dengan karakteristik Program Studi Sistem Informasi yang berfokus pada pengembangan solusi berbasis teknologi informasi untuk mendukung proses bisnis organisasi.

Distribusi anggota cluster pada penelitian ini menunjukkan kondisi yang tidak seimbang. Cluster 0 mencakup 353 dokumen atau sekitar 46,63% dari keseluruhan data, sedangkan Cluster 3 hanya terdiri dari 47 dokumen atau sekitar 6,21% dari total data. Ketidakseimbangan tersebut mengindikasikan bahwa minat penelitian mahasiswa tidak tersebar secara merata pada seluruh topik yang teridentifikasi. Beberapa tema penelitian cenderung lebih populer karena memiliki ruang implementasi yang luas, referensi yang melimpah, dan tingkat familiaritas yang tinggi bagi mahasiswa.

Selain merefleksikan kondisi nyata distribusi minat penelitian mahasiswa, ketidakseimbangan ukuran cluster juga dapat memengaruhi hasil evaluasi clustering. Cluster yang berukuran besar cenderung memiliki variasi dokumen yang lebih tinggi sehingga meningkatkan kemungkinan terjadinya overlap dengan cluster lain. Kondisi ini menjadi salah satu faktor yang menyebabkan nilai Silhouette Score pada penelitian relatif rendah. Meskipun demikian, hasil identifikasi kata dominan, visualisasi t-SNE, dan karakteristik kelayakan pada setiap cluster menunjukkan bahwa kelompok yang terbentuk masih memiliki interpretasi topik yang jelas dan relevan dengan konteks penelitian.

Hasil analisis kelayakan menunjukkan bahwa Cluster 4 yang didominasi topik monitoring, dashboard, dan IoT memiliki rata-rata skor kelayakan tertinggi sebesar 92,36. Selain itu, sebanyak 80% pengajuan pada cluster ini berstatus layak dan tidak ditemukan pengajuan yang berstatus ditolak. Temuan ini mengindikasikan bahwa topik berbasis monitoring dan Internet of Things masih memiliki peluang pengembangan yang luas sehingga dinilai memiliki tingkat kebaruan dan relevansi yang tinggi.

Sebaliknya, Cluster 5 yang berfokus pada tema wisata Danau Ranau dan sistem informasi pariwisata memiliki rata-rata skor kelayakan terendah sebesar 79,47 dengan persentase revisi mencapai 70,91%. Kondisi ini menunjukkan adanya kecenderungan pengulangan tema penelitian yang serupa sehingga tingkat kemiripan dengan penelitian sebelumnya relatif lebih tinggi. Temuan ini diperkuat oleh nilai similarity score rata-rata cluster tersebut yang mencapai 52,85, tertinggi dibandingkan cluster lainnya.

Cluster 3 yang didominasi topik Sistem Pendukung Keputusan (SPK) menggunakan metode SAW dan TOPSIS menunjukkan tingkat kelayakan yang cukup baik dengan persentase status layak sebesar 63,83%. Meskipun demikian, topik SPK merupakan salah satu bidang yang telah banyak diteliti sehingga mahasiswa perlu memperhatikan aspek kebaruan objek penelitian maupun metode yang digunakan agar tetap memenuhi kriteria kelayakan.

Pengujian menggunakan Kruskal-Wallis menghasilkan nilai statistik sebesar 49,813 dengan p-value $1{,}51\times10^{-9}$. Nilai tersebut lebih kecil dari tingkat signifikansi 0,05 sehingga

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

menunjukkan bahwa terdapat perbedaan skor kelayakan yang signifikan antar cluster. Hasil ini membuktikan bahwa topik penelitian memiliki hubungan yang signifikan terhadap tingkat kelayakan pengajuan judul skripsi mahasiswa. Dengan demikian, hasil clustering tidak hanya mampu mengelompokkan tema penelitian yang serupa, tetapi juga dapat digunakan untuk mengidentifikasi kelompok topik yang memiliki peluang kelayakan lebih tinggi dalam proses pengajuan judul skripsi.

Hasil penelitian ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa kombinasi TF-IDF dan K-Means efektif digunakan untuk mengidentifikasi pola topik pada kumpulan dokumen teks. Penelitian Siahaan dkk. [1] berhasil memetakan tren penelitian berdasarkan publikasi ilmiah, sedangkan Zendrato dkk. [6] memanfaatkan clustering untuk mengelompokkan dokumen tugas akhir mahasiswa berdasarkan tema penelitian. Kesamaan hasil tersebut menunjukkan bahwa TF-IDF dan K-Means tetap menjadi pendekatan yang efektif untuk analisis dokumen akademik.

Namun demikian, penelitian ini memiliki perbedaan dibandingkan penelitian terdahulu. Sebagian besar penelitian sebelumnya berfokus pada identifikasi topik atau pengelompokan dokumen tanpa mengaitkannya dengan indikator kualitas dokumen yang dianalisis. Pada penelitian ini, hasil clustering diintegrasikan dengan atribut similarity score, skor kelayakan, dan status pengajuan sehingga setiap cluster tidak hanya merepresentasikan topik penelitian, tetapi juga menunjukkan karakteristik tingkat kelayakan usulan judul skripsi. Hasil pengujian Kruskal–Wallis yang menunjukkan perbedaan signifikan skor kelayakan antar cluster memperkuat bahwa topik penelitian tertentu memiliki kecenderungan tingkat kelayakan yang berbeda. Temuan ini memberikan nilai tambah berupa pemahaman yang lebih komprehensif mengenai hubungan antara tema penelitian dan kualitas pengajuan judul skripsi mahasiswa.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan terhadap 757 judul skripsi unik mahasiswa Program Studi Sistem Informasi Fakultas Sains dan Teknologi UIN Raden Intan Lampung, dapat disimpulkan bahwa metode TF-IDF dan K-Means Clustering mampu digunakan untuk mengidentifikasi pola topik penelitian serta karakteristik kelayakan pengajuan judul skripsi. Tahapan preprocessing teks yang meliputi case folding, tokenisasi, stopword removal, dan stemming berhasil mengubah data judul menjadi representasi numerik yang dapat diproses menggunakan TF-IDF. Selanjutnya, proses clustering menghasilkan enam kelompok topik utama yang merepresentasikan kecenderungan penelitian mahasiswa, yaitu topik sistem informasi dan manajemen, aplikasi mobile dan layanan digital, sistem informasi desa dan pengelolaan data, sistem pendukung keputusan, sistem monitoring dan dashboard, serta sistem informasi pariwisata.

Hasil evaluasi menunjukkan bahwa nilai silhouette score sebesar 0,0314 mengindikasikan bahwa karakteristik data judul skripsi memiliki tingkat kemiripan yang cukup tinggi antar topik sehingga batas antar cluster tidak terbentuk secara tegas. Meskipun demikian, analisis kata kunci dominan dan contoh judul pada setiap cluster menunjukkan bahwa metode K-Means tetap mampu menghasilkan pengelompokan topik yang relevan dan mudah diinterpretasikan.

Analisis terhadap data kelayakan pengajuan judul menunjukkan adanya perbedaan karakteristik pada masing-masing cluster. Cluster sistem monitoring dan dashboard memiliki rata-rata skor kelayakan tertinggi sebesar 92,36 dengan persentase status layak mencapai 80%, sedangkan cluster pariwisata dan wisata Danau Ranau memiliki rata-rata skor kelayakan terendah sebesar 79,47 dengan dominasi status revisi sebesar 70,91%. Hasil uji Kruskal–Wallis menghasilkan nilai statistik sebesar 49,81 dengan p-value $1{,}51 \times 10^{-9}$, yang menunjukkan bahwa terdapat perbedaan signifikan skor kelayakan antar cluster. Temuan ini mengindikasikan bahwa topik penelitian memiliki hubungan dengan tingkat kelayakan pengajuan judul skripsi.

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

Penelitian ini memberikan manfaat bagi program studi dalam memetakan tren penelitian mahasiswa, mengevaluasi distribusi topik skripsi, serta mendukung proses pengambilan keputusan dalam penilaian usulan judul. Selain itu, hasil penelitian dapat menjadi dasar dalam penyusunan rekomendasi topik penelitian yang lebih relevan dan memiliki peluang kelayakan yang lebih tinggi.

Adapun keterbatasan penelitian ini terletak pada penggunaan data yang hanya berasal dari satu program studi serta pemanfaatan fitur berbasis TF-IDF yang belum mampu menangkap makna semantik secara mendalam. Untuk penelitian selanjutnya, disarankan menggunakan metode representasi teks berbasis embedding seperti Word2Vec, FastText, BERT, atau Sentence Transformer serta membandingkan performa algoritma clustering lain seperti Hierarchical Clustering, DBSCAN, dan Agglomerative Clustering. Selain itu, penelitian dapat diperluas dengan menambahkan variabel lain seperti bidang konsentrasi, dosen pembimbing, maupun lama penyelesaian skripsi untuk memperoleh pemahaman yang lebih komprehensif mengenai faktor-faktor yang memengaruhi kelayakan pengajuan judul skripsi.

5. REFERENCES

[1] J. P. Laksana, H. Irsyad, and A. Rahman, "Analisis Topik Dominan Dalam Paper Ilmu Komputer Menggunakan TF-IDF Dan K-Means," vol. 3, no. 3, pp. 78–84, 2025, doi: 10.58369/biit.v2i3.122.

[2] T. A. Br Sembiring and M. S. Hasibuan, "TEXT CLUSTERING IN KARO LANGUAGE USING TF-IDF WEIGHTING AND K-MEANS CLUSTERING," *J. Tek. Inform.*, vol. 4, no. 5 SE-Articles, pp. 1257–1265, Nov. 2023, doi: 10.52436/1.jutif.2023.4.5.1462.

[3] J. P. Pamput, A. R. Muthmainnah, A. Akram, N. Risal, and D. Fatmarani, "K-Means ++ and TF-IDF for Grouping Library Books by Topic," vol. 27, no. 2, pp. 74–82, 2025.

[4] M. Y. Hidayat, M. A. Yaqin, and Z. Abidin, "Semantic-Enhanced News Clustering Using TF-IDF and WordNet with K-Means," vol. 7, no. 4, pp. 3924–3951, 2025, doi: 10.63158/journalisi.v7i4.1260.

[5] Miquel Yosafat and Jatmika, "Implementasi Text Clustering Terkait Pilpres 2024 Menggunakan Metode K-Means," *Infact Int. J. Comput.* , vol. 8, no. 01 SE-Articles, pp. 6–12, Jan. 2024, doi: 10.61179/jurnalinfact.v8i01.496.

[6] F. S. Genius Zendrato, A. Triayudi, and E. T. E, "Analisis Clustering Dokumen Tugas Akhir Mahasiswa Sistem Informasi Universitas Nasional menggunakan Metode K-Means Clustering," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 6, no. 1 SE-Computer & Communication Science, pp. 70–76, 2022, doi: 10.35870/jtik.v6i1.389.

[7] K. Nurul, I. Djati, and N. Faiza, "Identifying Improvement Strategic from User Application Reviews Group Using K-Means Clustering and TF-IDF Weighting," vol. 7, no. 2, pp. 152–159, 2023.

[8] A. F. Zabidi, "Penerapan Algoritma K-Means untuk Pengelompokan Koleksi Perpustakaan dengan Data Mining," vol. 16, no. 2, 2024.

[9] J. ZHU, S. HUANG, Y. SHI, K. WU, and Y. WANG, "A Method of K-Means Clustering Based on TF-IDF for Software Requirements Documents Written in Chinese Language," *IEICE Trans. Inf. Syst.*, vol. E105.D, no. 4, pp. 736–754, 2022, doi: 10.1587/transinf.2021EDP7144.

[10] U. Buatoom, W. Kongprawechnon, and T. Theeramunkong, "Document Clustering Using K-Means with Term Weighting as Similarity-Based Constraints," 2020. doi: 10.3390/sym12060967.

[11] T. Bezdan *et al.*, "Hybrid Fruit-Fly Optimization Algorithm with K-Means for Text Document Clustering," 2021. doi: 10.3390/math9161929.

[12] M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," 2021. doi: 10.3390/e23060759.

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi

CHAIN: Journal of Computer Technology, Computer Engineering and Informatics
Website : https://ejournal.techcart-press.com/index.php/chain
E-ISSN 2964-2485
P-ISSN 2964-2450
Volume 4, Number 3, June 2026
DOI : **https://doi.org/10.58602/chain.v4i3.296**
Page 270-283

[13] D. Chicco, A. Sichenze, and G. Jurman, *A simple guide to the use of Student ′ s t - test , Mann - Whitney U test , Chi - squared test , and Kruskal - Wallis test in biostatistics*. BioMed Central, 2025.

*Singgih Yulizar Ma'ruf: *Penulis Korespondensi