JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan Volume 3, No. 4, Tahun 2026
ISSN 3025-0927 (media online)
Hal 445-453

Model Hybrid Clustering untuk Artikel Ilmiah Prosiding SENATIB Berbasis TF-IDF

Nabila Wahyu Ningtias[1], Abdullah Azza Al Abllas[2*], Arif Kurniawan[3], Nasywa Hamna Zakiya[4], Aprilisa Arum[5]

[1-5]Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Duta Bangsa, Surakarta, Indonesia
Email: [1]nabilawahyu.id@gmail.com, [2]azzul7665@gmail.com*, [3]arifkurniawan47203@gmail.com, [4]nsywahmna@gmail.com, [5]aprilisa_arumsari@udb.ac.id
(* : coressponding author)

Abstrak−Meningkatnya jumlah artikel ilmiah pada prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB) menyebabkan proses pencarian referensi berdasarkan topik penelitian menjadi kurang efisien apabila dilakukan secara manual. Penelitian ini bertujuan mengelompokkan artikel ilmiah berdasarkan kemiripan topik menggunakan pendekatan text mining dan algoritma K-Means Clustering. Data penelitian diperoleh melalui proses web scraping sebanyak 490 artikel yang terdiri atas atribut judul dan abstrak. Tahapan penelitian meliputi data cleaning, preprocessing (case folding, cleaning, tokenization, stopword removal), pembobotan menggunakan Term Frequency–Inverse Document Frequency (TF-IDF), penentuan jumlah cluster menggunakan Agglomerative Hierarchical Clustering dengan Ward Linkage dan Elbow Method, dilanjutkan proses clustering menggunakan K-Means. Hasil penelitian menunjukkan bahwa jumlah cluster yang paling representatif adalah tiga cluster. Evaluasi menggunakan Silhouette Score, Davies–Bouldin Index (DBI), dan Calinski–Harabasz Index (CHI) menghasilkan nilai masing-masing sebesar 0,0336, 4,5271, dan 13,7372. Meskipun tingkat pemisahan antarcluster masih relatif rendah akibat kemiripan topik antarartikel, hasil clustering mampu mengelompokkan artikel ke dalam tiga kelompok utama, yaitu Sistem Informasi, Internet of Things (IoT), dan Sistem Pendukung Keputusan. Penelitian ini diharapkan dapat membantu proses pencarian referensi berdasarkan topik serta memberikan gambaran mengenai persebaran topik penelitian pada prosiding SENATIB.

Kata Kunci: text mining; TF-IDF; K-Means Clustering; Agglomerative Hierarchical Clustering; SENATIB.

Abstract−The increasing number of scientific articles published in the Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB) proceedings has made the process of searching for references based on research topics less efficient when performed manually. This study aims to cluster scientific articles based on topic similarity using a text mining approach and the K-Means Clustering algorithm. The dataset was collected through a web scraping process, resulting in 490 articles consisting of title and abstract attributes. The research stages included data cleaning, text preprocessing (case folding, cleaning, tokenization, and stopword removal), feature weighting using Term Frequency–Inverse Document Frequency (TF-IDF), determining the optimal number of clusters using Agglomerative Hierarchical Clustering with the Ward Linkage method and the Elbow Method, followed by document clustering using the K-Means algorithm. The results indicated that the most representative number of clusters was three. Cluster quality evaluation using the Silhouette Score, Davies–Bouldin Index (DBI), and Calinski–Harabasz Index (CHI) produced values of 0.0336, 4.5271, and 13.7372, respectively. Although the cluster separation was relatively low due to the similarity of research topics among articles, the clustering process successfully grouped the documents into three main categories: Information Systems, Internet of Things (IoT), and Decision Support Systems. The proposed approach is expected to facilitate topic-based reference retrieval and provide insights into the distribution of research topics in the SENATIB proceedings.

Keywords: text mining; TF-IDF; K-Means Clustering; Agglomerative Hierarchical Clustering; SENATIB.

1. PENDAHULUAN

Meningkatnya jumlah publikasi ilmiah di berbagai bidang, khususnya informatika, sistem informasi, Artificial Intelligence, machine learning, Internet of Things (IoT), dan sistem pendukung keputusan adalah hasil dari kemajuan teknologi informasi. Setiap tahun, prosiding SENATIB, yang diselenggarakan secara teratur, memuat ratusan artikel tentang berbagai topik penelitian. Banyak artikel memberikan kemudahan dalam memperoleh referensi, tetapi proses pencarian dan pengenalan artikel yang sesuai dengan toik penelitian juga sulit. Melakukan pencarian dengan kata kunci atau membaca judul dan abstrak secara manual memerlukan waktu yang cukup lama dan tidak dapat memberikan pengelompokan topik secara otomatis. Oleh karena itu, untuk menjadikan proses

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan Volume 3, No. 4, Tahun 2026
ISSN 3025-0927 (media online)
Hal 445-453

pencarian referensi dan analisis perkembangan topik penelitian lebih efisien, diperlukan suatu pendekatan yang dapat mengelompokkan artikel berdasarkan tingkat kemiripan isi(Prasetya & Sulistiani, 2023).

Text mining adalah teknik pengolahan data teks yang mengekstraksi informasi melalui tahapan preprocessing, seperti perbaikan, pembongkaran case, tokenisasi, enghapusan stopword, dan pembobotan kata menggunakan Frekuensi Term-Inverse Dokumen (TF-IDF). Salah satu metode yang dapat digunakan untuk mengatasi masalah ini. Judul dan abstrak artikel ilmiah dapat menggambarkan isi penelitian, sehingga dapat digunakan sebagai sumber data selama proses analisis. Clustering dokumen adalah metode text mining yang tidak diawasi pembelajaran yang mengelompokkan dokumen berdasarkan tingkat kemiripan atribut. Metode ini tidak membutuhkan label kelas untuk mengelompokkan sejumlah besar dokumen, K-Means adalah algoritma clustering yang palling populer. Namun demikian, proses preprocessing, representasi fitur menggunakan TF-IDF, dan ketepatan dalam menghitung jumlah cluster sangat memengaruhi kualitas hasil clustering (Wahyudi & Nugroho, 2022);(Ikotun et al., 2023).

TF-IDF dan algoritma K-Means telah digunakan dalam proses clustering dokumen dalam beberapa penelitian. Studi Widaningrum dkk. (2022) mengelompokkan 93 dokumen penelitian dengan TF-IDF dann K-Means. Mereka juga menggunakan Silhouette Score untuk menentukan jumlah cluster yang tepat. Hasilnya adalah empat cluster yanng ideal. Dengan tokenisasi, case folding, dan removal of stopwords, Sembiring dan Hasibuan (2023) memanfaatkan TF-IDF dan K-Means untuk mengelompokkan dokumen bahasa Karo berdasarkan kemiripan dialek. Surianto (2025) juga mengembangkan proses clustering artikel ilmiah dengan manggabungkan TF-IDF dan Latent Dirichlet Allocation (LDA) sebagai ekstraksi fitur sebelum proses K-Means, yang menghasilkan kualitas clustering yang lebih baik. Menurut penelitian tambahan yang dilakukan oleh Akbar dkk. (2025), metode pembelajaran yang tidak diawasi memiliki kemampuan untuk mengelopokkan dokumen ilmiah berdasarkan tingkat kemiripan teks, yang membuat proses analisis topik penelitian menjadi lebih mudah. Hasil penelitian menunjukkan bahwa algoritma K-Means masih banyak digunakan untuk memcluster dokumen karena sangat baik untuk data teks (Widaningrum et al., 2022);(Sembiring & Hasibuan, 2023).

Namun, penelitian sebelumnya tidak melakukan analisis struktur data alami menggunakan pendekatan hierarkis sebelum menerapkan TF-IDF dan algoritma K-Means. Selain itu, evaluasi hasil clustering biasanya hanya menggunakan satu teknik, seperti Silhouette Score, sehingga kualitas cluster belum dievaluasi secara menyeluruh. Selain itu, penelitian yang dilakukan pada artikel ilmiah dalam publikasi SENATIB masih sangat terbatas, meskipun publikasi tersebut mencakup berbagai topik penelitian yang dapat digunakan untuk mengetahui tren penelitian di bidang teknologi informasi (Surianto, 2025);(Sembiring & Hasibuan, 2023).

Tabel 1. Perbandingan Penelitian Terdahulu

Peneliti	Objek Penelitian	Metode	Evaluasi	Keterbatasan
Widaningrum dkk. (2022)	93 dokumen penelitian	TF-IDF + K-Means	Silhouette Score	Belum menggunakan analisis hierarki dan evaluasi multi-metrik
Sembiring & Hasibuan (2023)	Dokumen bahasa Karo	TF-IDF + K-Means	Hasil pembentukan cluster	Objek penelitian bukan artikel ilmiah dan belum mengevaluasi kualitas cluster secara kuantitatif
Surianto (2025)	427 abstrak artikel jurnal	TF-IDF + LDA + K-Means	Silhouette Score	Berfokus pada peningkatan fitur menggunakan LDA, belum menggunakan

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan Volume 3, No. 4, Tahun 2026
ISSN 3025-0927 (media online)
Hal 445-453

				pendekatan hierarkis
Akbar dkk. (2025)	Artikel ilmiah	*Unsupervised Similarity-Based Clustering*	Evaluasi *unsupervised clustering*	Belum mengombinasikan analisis hierarki, Elbow Method, dan evaluasi multi-metrik
Penelitian ini	490 artikel prosiding SENATIB	Web Scraping – TF-IDF – Agglomerative Hierarchical Clustering – Dendrogram – Elbow Method – K-Means – PCA	Silhouette Score, Davies–Bouldin Index, Calinski–Harabasz Index	Menyediakan proses ***clustering*** yang lebih komprehensif mulai dari pengumpulan data hingga evaluasi hasil

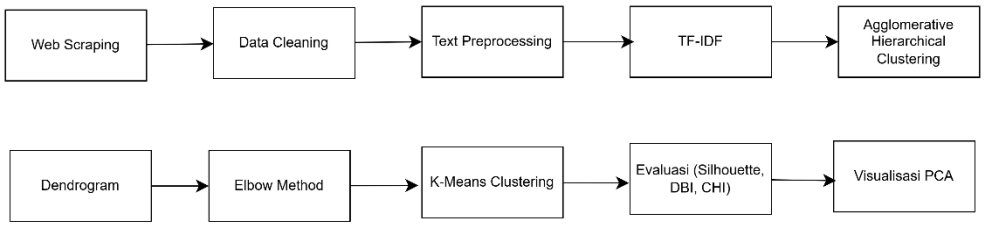
Penelitian ini mengusulkan metode pengelompokkan artikel ilmiah yang lebih komprehensif yang menggabungkan Agglomerative Hierarchical Clustering dengan Linkage Ward untuk memperoleh struktur awal data melalui visualisasi dendrogram. Metode ini didasarkan pada masalah dan penelitian terdahulu. Sebelum memulai proses pengelompokkan menggunakan algoritma K-Means, metode Elbow digunakan untuk memastikan jumlah cluster. Selanjutnya, hasil clustering dievaluasi menggunakan Silhouette Score, Davies–Bouldin Index (DBI), dan Calinski–Harabasz Index (CHI). Untuk mempermudah interpretasi distribusi antarcluster, Principal Component Analysis (PCA) juga digunakan.

Untuk penelitian ini, 490 artikel prosiding SENATIB diperoleh melalui scraping web. Nama penulis, judul artikel, dan abstrak adalah komponen data yang dikumpulkan, tetapi proses analisis hanya menggunakan judul dan abstrak karena keduanya lebih baik menunjukkan topik penelitian. Hasil penelitian diharapkan dapat memberikan gambaran tentann bagaimana topik penelitian tersebar di prosiding SENATIB, membantu peneliti menemukan referensi yang sesuai dengan bidang mereka, dan menjadi referensi untuk pengembangan penelitian clustering dokumen ilmiah dan text mining.

2. METODE PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menggunakan pendekatan text mining untuk mengelompokkan artikel ilmiah prosiding SENATIB berdasarkan kemiripan topik penelitian. Tahapan penelitian dimulai dari proses pengumpulan data melalui web scraping, dilanjutkan dengan preprocessing data teks, pembobotan kata menggunakan TF-IDF, penentuan jumlah cluster menggunakan Agglomerative Hierarchical Clustering dan Elbow Method, proses pengelompokan menggunakan algoritma K-Means, hingga evaluasi hasil clustering. Alur penelitian ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan Volume 3, No. 4, Tahun 2026
ISSN 3025-0927 (media online)
Hal 445-453

2.2 Pengumpulan Data

Data penelitian dikumpulkan melalui scrapiing web dari artiikel prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB). Scraping membuat data lebih lengkap dan efisien daripada pengumpulan data manual. Data yang berhasil dikumpulkan terdiri dari 490 artikel yang memiliki atribut nama penullis, judul dan abstrak. Aspek judul dan abstrak digunakan sebagai sumber utama dalam proses analisis penelitian karena atribut-atribut ini lebih baik mewakili isi penelitian daripada atribut lainnya.

2.3 Data Cleaning

Sebelum proses analisis dimulai, langkah pembersihan data dilakukan untuk memastikan kualitas data. Pada titik ini, semua atribut diperiksa untuk kesesuaian format, data duplikat, dan data yang tidak memiliki nilai (missing value). Untuk menghindari pengaruh pada hasil clustering, artikel yang mengandung informasi yang tidak lengkap atau terduplikasi dihapus. Selain itu, pemeriksaan panjang abstrak dilakukan untuk memastikan bahwa setiap dokumen mengandung informasi yang cukup menggambarkan isi artikel.

2.3.1 Case Folding

Untuk mencegah perbedaan penulisan antara huruf kapital dan huruf kecil, tahap case folding digunakan untuk mengubah semua huruf dalam judul dan abstrak menjadi huruf kecil (lowercase). Sebagai contoh pada penelitian ini, judul "Perancangan Sistem Informasi Penjualan Berbasis Web" diubah menjadi "perancangan sistem informasi penjualan berbasis web".

2.3.2 Cleaning

Pada penelitian ini, tahap cleaning dilakukan dengan menghapus karakter yang tidak diperlukan, seperti angka, tanda baca, simbol, URL, dan karakter khusus lainnya. Sebagai contoh pada penelitian ini, teks "Perancangan Sistem Informasi Berbasis Web Tahun 2024!!!" diubah menjadi "perancangan sistem informasi berbasis web tahun"

2.3.3 Tokenizing

Pada penelitian ini, proses tokenization dilakukan dengan memecah setiap kalimat menjadi kumpulan kata sehingga setiap kata dapat diproses secara terpisah. Sebagai contoh pada penelitian ini, kalimat "perancangan sistem informasi berbasis web tahun" diubah menjadi ['perancangan', 'sistem', 'informasi', 'berbasis', 'web', 'tahun']

2.3.4 Stopword Removal

Penelitian ini menggunakan stopword removal untuk menghilangkan kata-kata umum yang tidak memberikan informasi penting tentagn subjek penelitian. Penelitian ini tidak hanya menggunakan daftar stopword Bahasa Indonesia, tetapi juga menambahkan beberapa stopword khusus yang sering muncul dalam abstrak artikel. Namun, ini tidak membantu membedakan topik penelitian. Sebagai contoh pada penelitian ini, token ['penelitian', 'ini', 'menggunakan', 'metode', 'kmeans', 'clustering'] diubah menjadi ['kmeans', 'clustering']

2.3.5 Pembentukan Clean Text

Pada penelitian ini, token hasil preprocessing kemudian digabungkan kembali menjadi satu dokumen (clean text) yang digunakan sebagai masukan pada proses pembobotan TF-IDF. Sebagai contoh pada penelitian ini, token ['sistem', 'informasi', 'penjualan', 'web'] digabungkan menjadi "sistem informasi penjualan web"

2.4 Pembobotan TF-IDF

Setelah proses preprocessing selesai, metode *Term Frequency–Inverse Document Frequency* (TF-IDF) digunakan untuk mengubah setiap dokumen menjadi representasi numerik. Metode ini memberikan bobot berdasarkan seberapa penting suatu kata terhadap dokumen secara keseluruhan. Kata-kata yang sering muncul di satu dokumen tetapi jarang muncul di dokumen lain akan diberi bobot yang lebih tinggi, sehingga mereka dapat mewakili topik penelitian

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan Volume 3, No. 4, Tahun 2026
ISSN 3025-0927 (media online)
Hal 445-453

dengan lebih baik. Selanjutnya, hasil pembobotan TF-IDF digunakan sebagai input untuk proses clustering(Mehta et al., 2024).

2.5 Penentuan Jumlah Cluster

Penelitian ini menggunakan dua metode untuk menghitung jumlah cluster sebelum proses K-Means. Metode awal menggunakan Agglomerative Hierarchical Clustering melalui pendekatan Ward Linkage. Untuk menggambarkan struktur alami data, hasil pengelompokkan divisualisasikan dalam bentuk dendrogram. Dendrogram ini dapat digunakan sebagai referensi untuk menentukan jumlah cluster yang paling representatif (Wicaksono et al., 2024).
Metode Elbow digunakan dalam pendekatan kedua. Dalam metode ini, nili Sum of Squared Error (SSE) dihitung untuk sejumlah jumlah cluster tertentu. Nilai SSE yang membentuk titik siku (elbow) digunakan sebagai perbandingnan terhadap hasil analisis dendrogram sebelum jumlah cluster akhir ditetapkan (Ikotun et al., 2023).

2.6 K-Means Clustering

Algoritma pengelompokan K-Means digunakan untuk melakukan proses pengelompokan dokumen. Algoritma ini membagi dokumen ke dalam berbagai cluster berdasarkan jarak dari centroid (pusat cluster). Iterasi dilakukan hingga posisi centroid tidak berubah atau stabil. Hasil akhir adalah Kumpulan artikel yang memiliki Tingkat kemiripan topik penelitian (Ikotun et al., 2023).

2.7 Evaluasi Clustering

Kualitas hasil clustering dievaluasi menggunakan tiga metrik, yaitu:

a. Silhouette Score, untuk mengukur tingkat kemiripan suatu dokumen terhadap cluster tempatnya berada dibandingkan dengan cluster lain(Ikotun et al., 2025).
b. Davies–Bouldin Index (DBI), untuk mengukur tingkat pemisahan antarcluster. Nilai DBI yang lebih kecil menunjukkan kualitas clustering yang lebih baik(Ikotun et al., 2025).
c. Calinski–Harabasz Index (CHI), untuk mengukur rasio antara variasi antarcluster dan variasi di dalam cluster. Nilai CHI yang lebih besar menunjukkan hasil clustering yang semakin baik(Hassan, 2024).

Penggunaan ketiga metrik tersebut bertujuan memberikan evaluasi yang lebih komprehensif dibandingkan hanya menggunakan satu metode evaluasi(Dewi et al., 2024).

2.8 Visualisasi Cluster

Hasil clustering divisualisasikan menggunakan Principal Component Analysis (PCA). PCA digunakan untuk mengurangi dimensi data TF-IDF menjadi dua dimensi, sehingga distribusi setiap cluster dapat ditampilkan dalam bentuk grafik. Visualisasi ini memberikan Gambaran tentang pola persebaran dokumen hasil clustering dan membantu mengamati Tingkat pemisahan antar cluster (Rosyada et al., 2024).

3. ANALISA DAN PEMBAHASAN

3.1 Hasil Pengumpulan Data

Pada penelitian ini, data diperoleh melalui proses web scraping terhadap artikel prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB). Proses scraping menghasilkan 490 artikel ilmiah yang terdiri atas atribut nama penulis, judul artikel, dan abstrak. Atribut judul dan abstrak digunakan sebagai sumber utama dalam proses text mining karena mampu merepresentasikan topik penelitian secara ringkas.

Tabel 2. Jumlah Dataset

Atribut	Jumlah
Jumlah artikel	490
Nama penulis	490
Judul	490

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan Volume 3, No. 4, Tahun 2026
ISSN 3025-0927 (media online)
Hal 445-453

Abstrak	490

Setelah data berhasil dikumpulkan, dilakukan proses data cleaning untuk memastikan tidak terdapat data duplikat maupun data yang kosong. Berdasarkan hasil pemeriksaan, seluruh dokumen memenuhi kriteria sehingga dapat digunakan pada tahap analisis berikutnya.

3.2 Hasil Preprocessing

Tahap preprocessing dilakukan untuk menghasilkan data teks yang lebih bersih sebelum proses pembobotan TF-IDF. Tahapan yang dilakukan meliputi case folding, cleaning, tokenization, stopword removal, dan pembentukan clean text. Setiap tahapan bertujuan menghilangkan informasi yang tidak relevan sehingga kata-kata yang tersisa lebih mampu merepresentasikan isi dokumen.

Tabel 3. Hasil Preprocessing

Tahapan	Hasil
Judul	Faktor-Faktor yang Mempengaruhi Pelaksanaan Green Banking...
Abstrak	Berbagai isu yang menjadi prioritas utama lembaga...
Tokenisasi	faktorfaktor mempengaruhi pelaksanaan green banking berbagai isu menjadi prioritas utama ...
Stopword Removal	faktorfaktor mempengaruhi pelaksanaan green banking prioritas utama lembaga ...
Clean Text	faktorfaktor mempengaruhi pelaksanaan green banking prioritas utama lembaga ...

Berdasarkan hasil preprocessing, kata-kata yang tidak memiliki makna penting berhasil dihilangkan sehingga dokumen menjadi lebih ringkas dan siap digunakan pada proses pembobotan TF-IDF.

3.3 Pembobotan TF-IDF

Setelah proses preprocessing, setiap dokumen diubah menjadi representasi numerik menggunakan metode TF-IDF. Hasil pembobotan menghasilkan matriks berukuran 490 × 500. Artinya terdapat 490 dokumen dan 500 fitur yang digunakan sebagai representasi kata pada proses clustering. Penggunaan batas maksimum 500 fitur bertujuan mengurangi dimensi data sekaligus mempertahankan kata-kata yang paling informatif.

$$TFIDF(t,d) = TF(t,d) \times IDF(t)$$

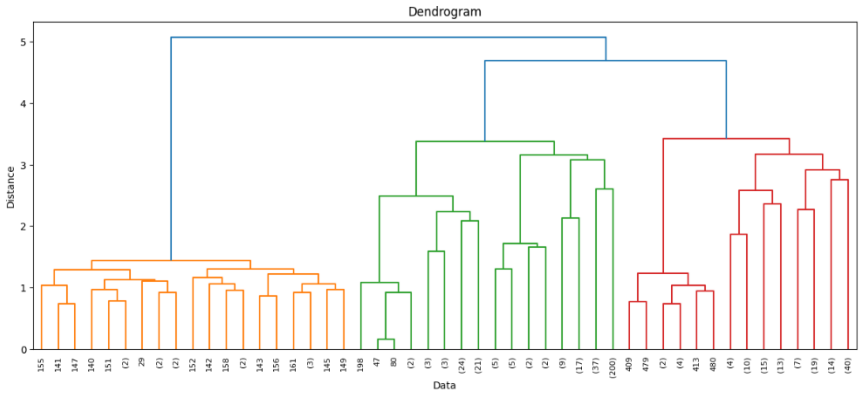
3.4 Penentuan Jumlah Cluster

Sebelum dilakukan proses K-Means, penelitian ini menentukan jumlah cluster menggunakan dua pendekatan, yaitu Agglomerative Hierarchical Clustering dan Elbow Method.

3.4.1 Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering diterapkan menggunakan metode Ward Linkage untuk mengetahui struktur alami data. Hasil pengelompokan divisualisasikan menggunakan dendrogram seperti pada Gambar 2.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan Volume 3, No. 4, Tahun 2026
ISSN 3025-0927 (media online)
Hal 445-453

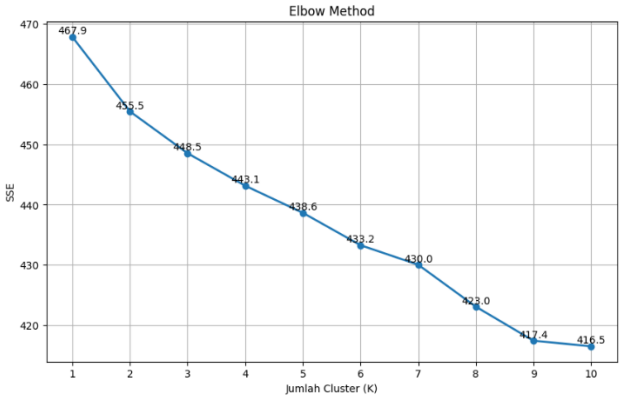


Gambar 2. Dendrogram Agglomerative Hierarchical Clustering

Berdasarkan dendrogram, terlihat bahwa pemotongan pada tingkat jarak tertentu menghasilkan tiga kelompok utama, sehingga jumlah cluster awal ditetapkan sebanyak 3 cluster. Penggunaan metode Ward dipilih karena mampu meminimalkan variasi di dalam cluster sehingga menghasilkan kelompok yang relatif lebih homogen.

3.4.2 Elbow Method

Selanjutnya dilakukan pengujian menggunakan Elbow Method dengan menghitung nilai Sum of Squared Error (SSE) pada jumlah cluster 1 sampai 10.



Gambar 3. Grafik Elbow Method

Berdasarkan grafik tersebut, penurunan nilai SSE mulai melambat setelah K = 3, sehingga jumlah cluster tersebut dianggap cukup representatif. Hasil ini juga sejalan dengan analisis menggunakan dendrogram yang menunjukkan adanya tiga kelompok utama.

3.5 Hasil K-Means Clustering

Berdasarkan hasil penentuan jumlah cluster, algoritma K-Means diterapkan menggunakan K = 3. Proses ini menghasilkan tiga kelompok artikel berdasarkan tingkat kemiripan isi judul dan abstrak. Distribusi dokumen pada setiap cluster ditunjukkan pada Tabel 4.

Tabel 4. Jumlah Artikel Tiap Cluster

Cluster	Jumlah cluster
Cluster 0	303
Cluster 1	135
Cluster 2	52

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan Volume 3, No. 4, Tahun 2026
ISSN 3025-0927 (media online)
Hal 445-453

3.6 Evaluasi Clustering

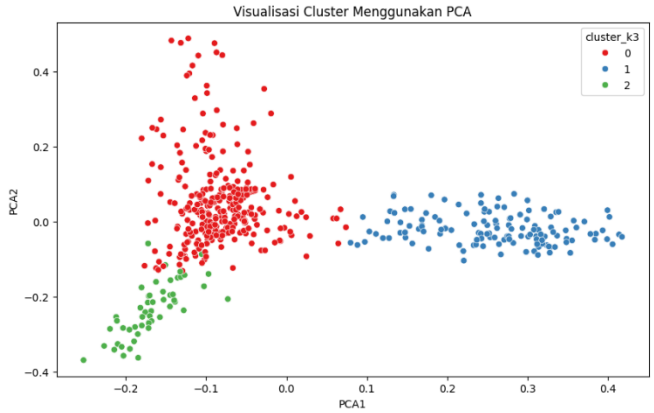
Tabel 5. Evaluasi Clustering

Metrik	Nilai
Silhouette Score	0.0336
Davies–Bouldin Index	4.5271
Calinski–Harabasz Index	13.7372

Sebagai hasil dari nilai Silhouette Score, Tingkat pemisahan antar cluster masih sangat rendah. Ini disebabkan oleh fakta bahwa beberapa dokumen memiliki karakteristik yang serupa karena artikel ilmiah dala prosiding SENATIB memiliki topik yang sangat terkait. Sementara nilai indeks Calinski-Harabasz menunjukkan bahwa struktur cluster masih dapat membedakan kelompok dokumen berdasarkan karakteristik utamanya, nilai indeks Davies-Bouldin menunjukkan bahwa pemisahan antarcluster belum sepenuhnya ideal. Meskipun demikian, hsil interpretasi kata dominan menunjukkan bahwa setiap kelompok berhasil menggambarkan subjek penelitian yang berbeda. SIstem Informasi, Internet pf Things (IoT), da Sistem Pendukung Keputusan adalah contoh dari kelompok ini.

3.7 Visualisasi Cluster

Untuk mempermudah interpretasi hasil clustering, dilakukan visualisasi menggunakan Principal Component Analysis (PCA).



Gambar 4. Visualisasi Hasil Clustering Menggunakan PCA

Visualisasi menunjukkan bahwa sebagian besar dokumen berhasil membentuk tiga kelompok utama sesuai hasil pengelompokan K-Means. Meskipun masih terdapat beberapa titik yang saling berdekatan akibat kemiripan antar topik penelitian, secara umum distribusi dokumen telah menunjukkan pola pengelompokan yang sesuai dengan karakteristik masing-masing cluster.

4. KESIMPULAN

Penelitian ini berhasil menerapkan pendekatan text mining untuk mengelompokkan 490 artikel prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB) menggunakan pembobotan Term Frequency–Inverse Document Frequency (TF-IDF) dan algoritma K-Means Clustering. Sebelum proses clustering, dilakukan penentuan jumlah cluster melalui Agglomerative Hierarchical Clustering dengan metode Ward Linkage dan Elbow Method, yang menunjukkan bahwa jumlah cluster yang paling representatif adalah tiga cluster. Hasil evaluasi menggunakan Silhouette Score sebesar 0,0336, Davies–Bouldin Index (DBI) sebesar 4,5271, dan Calinski–Harabasz Index (CHI) sebesar 13,7372 menunjukkan bahwa meskipun pemisahan antarcluster belum optimal akibat tingginya kemiripan topik antarartikel, proses clustering telah mampu

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan Volume 3, No. 4, Tahun 2026
ISSN 3025-0927 (media online)
Hal 445-453

mengelompokkan artikel ke dalam tiga kelompok utama, yaitu Sistem Informasi, Internet of Things (IoT), dan Sistem Pendukung Keputusan. Dengan demikian, pendekatan yang diusulkan dapat membantu proses pengelompokan artikel ilmiah secara otomatis serta mempermudah pencarian referensi berdasarkan topik penelitian. Untuk penelitian selanjutnya, disarankan menggunakan metode representasi teks yang lebih mampu menangkap hubungan semantik, seperti Word2Vec, FastText, Doc2Vec, atau BERT, serta membandingkan performa algoritma clustering lain seperti DBSCAN, Spectral Clustering, atau Hierarchical Clustering agar kualitas pengelompokan dokumen dapat ditingkatkan.

REFERENCES

Dewi, L. P., Sukarno, P., & Nugroho, A. S. (2024). Evaluasi metrik clustering pada dokumen teks berbahasa Indonesia menggunakan Silhouette Score, DBI, dan CHI. Jurnal Teknologi Dan Sistem Komputer, 12(3), 118–127. https://doi.org/10.14710/jtsiskom.v12i3.4892

Hassan, B. A. (2024). From A-to-Z Review of Clustering Validation Indices. ArXiv.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences, 622, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E. (2025). Cluster validity indices for automatic clustering: A comprehensive review. Heliyon.

Mehta, V., Agarwal, M., & Kaliyar, R. (2024). A comprehensive and analytical review of text clustering techniques. International Journal of Data Science and Analytics. https://doi.org/10.1007/s41060-024-00540-x

Prasetya, D. A., & Sulistiani, H. (2023). Penerapan text mining untuk analisis kecenderungan topik penelitian pada prosiding ilmiah. Jurnal Sistem Dan Teknologi Informasi, 11(4), 312–321. https://doi.org/10.26418/justin.v11i4.64187

Rosyada, I. A., Utari, D. T., Rosyada, I. A., & Utari, D. T. (2024). Penerapan Principal Component Analysis untuk Reduksi Variabel pada Algoritma K-Means Clustering Penerapan Principal Component Analysis untuk Reduksi Variabel pada Algoritma K-Means Clustering. 5(1), 6–13.

Sembiring, R. K., & Hasibuan, J. S. (2023). Pengelompokan dokumen berbahasa Karo menggunakan TF-IDF dan K-Means Clustering. JURTEKSI, 9(3), 201–210. https://doi.org/10.33330/jurteksi.v9i3.2134

Surianto, D. A. (2025). Kombinasi TF-IDF dan Latent Dirichlet Allocation untuk peningkatan kualitas clustering artikel ilmiah. Jurnal Sistem Informasi Dan Teknologi, 7(1), 33–44. https://doi.org/10.37034/jsisfotek.v7i1.305

Wahyudi, A., & Nugroho, H. A. (2022). Implementasi TF-IDF untuk representasi fitur teks pada pengelompokan dokumen ilmiah. JATISI, 9(2), 756–768. https://doi.org/10.35957/jatisi.v9i2.2143

Wicaksono, A. F., Hidayat, R., & Sari, D. K. (2024). Document clustering menggunakan TF-IDF berbasis web scraping pada prosiding seminar nasional. JATISI, 11(2), 445–458. https://doi.org/10.35957/jatisi.v11i2.7218

Widaningrum, I., Rizqiana, U., & Rahmat, A. (2022). Pengelompokan dokumen penelitian menggunakan TF-IDF dan K-Means Clustering. Jurnal Ilmiah Teknologi Informasi, 12(2), 145–155. https://doi.org/10.28932/jiti.v12i2.5832