Buletin Ilmiah Informatika Teknologi
Vol 3, No 3, Mei 2025, Hal. 78 - 84
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.122
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

Analisis Topik Dominan Dalam Paper Ilmu Komputer Menggunakan TF-IDF Dan K-Means

Jovansa Putra Laksana, Shela, Hafiz Irsyad, Abdul Rahman

[1]Fakultas Ilmu Komputer dan Rekayasa, Program Studi Informatika, Universitas Multi Data Palembang, Kota Palembang, Indonesia
Email: [1]jovansaputralaksana_2226250050@mhs.mdp.ac.id, [2]shela_2226250087@mhs.mdp.ac.id, [3]hafizirsyad@mdp.ac.id, [4]arahman@mdp.ac.id
Email Penulis Korespondensi: shela_2226250087@mhs.mdp.ac.id

Abstrak−Pertumbuhan jumlah publikasi ilmiah di bidang ilmu komputer mendorong kebutuhan untuk memahami distribusi dan tren topik yang berkembang. Penelitian ini bertujuan untuk mengidentifikasi dan menganalisis topik-topik dominan dalam publikasi tersebut menggunakan pendekatan *text mining* dengan *vektorisasi Term Frequency–Inverse Document Frequency* (TF-IDF) serta algoritma *clustering K-Means*. Data berupa 1.222 judul publikasi dari Semantic Scholar (2020–2025) dianalisis melalui tahapan normalisasi bahasa, pra-pemrosesan teks, ekstraksi fitur TF-IDF, penentuan jumlah klaster optimal, serta evaluasi kualitas klaster menggunakan *Silhouette Score* dan *Davies-Bouldin Index* (DBI). Hasil penelitian menunjukkan bahwa topik seperti keamanan siber, kecerdasan buatan, dan pembelajaran mesin merupakan tema yang dominan. Meskipun beberapa klaster menunjukkan kohesi internal yang baik, evaluasi global menghasilkan nilai *Silhouette Score* sebesar 0,0585 dan DBI sebesar 4,387, yang mengindikasikan adanya tumpang tindih antar klaster dan kurangnya pemisahan topik yang jelas. Temuan ini menunjukkan bahwa metode berbasis TF-IDF dan K-Means memiliki keterbatasan dalam menangkap konteks semantik, sehingga diperlukan pengembangan pendekatan representasi dan pengelompokan yang lebih kontekstual untuk meningkatkan kualitas analisis topik ke depan.

Kata Kunci: Analisis Topik; *Davies-Bouldin Index*; *K-Means*; Penambangan Teks; *Silhouette Score*; TF-IDF

Abstract−*The rapid growth of scientific publications in the field of computer science has created a need to understand the distribution and trends of emerging research topics. This study aims to identify and analyze dominant topics in computer science literature using a text mining approach based on Term Frequency–Inverse Document Frequency (TF-IDF) vectorization and the K-Means clustering algorithm. A total of 1,222 publication titles from Semantic Scholar (2020–2025) were processed through language normalization, text preprocessing, TF-IDF feature extraction, optimal cluster determination, and cluster quality evaluation using Silhouette Score and Davies-Bouldin Index (DBI). The results reveal that topics such as cybersecurity, artificial intelligence, and machine learning are the most prevalent. While some clusters show good internal cohesion, the overall evaluation yielded a Silhouette Score of 0.0585 and a DBI of 4.387, indicating overlapping topics and limited cluster separation. These findings suggest that although the TF-IDF and K-Means approach can highlight general topic trends, it has limitations in capturing semantic context. Future research is encouraged to explore more contextual representation and clustering techniques to improve topic analysis quality.*

Keywords: *Davies-Bouldin Index*; *K-Means*; *Silhouette Score*; *Text Mining*; TF-IDF; *Topic Analysis*

1. PENDAHULUAN

Perkembangan ilmu komputer yang begitu pesat telah mendorong peningkatan jumlah publikasi ilmiah secara signifikan setiap tahunnya. Lonjakan ini menghadirkan tantangan tersendiri bagi institusi akademik dan para peneliti dalam mengidentifikasi tren serta topik penelitian yang sedang berkembang. Untuk menjawab tantangan tersebut, pendekatan *text mining* menjadi solusi yang efektif dalam menganalisis dan mengelompokkan dokumen berdasarkan kemiripan topik. Sebagai ilustrasi, metode *text mining* menggunakan algoritma *K-Means* dan *cosine similarity* telah diterapkan untuk mengelompokkan dokumen penelitian dosen, yang menghasilkan enam klaster topik utama [1]. Studi lain menunjukkan efektivitas algoritma *K-Means* dalam mengelompokkan data survei berdasarkan kesamaan topik [2]. Sementara itu, metode *text mining* dan *K-Means* clustering juga telah dimanfaatkan untuk mengelompokkan dokumen skripsi mahasiswa guna mengungkap pola dan keterkaitan antar penelitian [3].

Salah satu teknik yang banyak digunakan dalam *text mining* adalah *Term Frequency-Inverse Document Frequency* (TF-IDF), yang berperan dalam mengekstraksi serta merepresentasikan fitur penting dari suatu teks. Teknik ini telah terbukti efektif dalam berbagai penelitian, khususnya dalam proses klasifikasi dan pengelompokan dokumen. Misalnya, penggabungan TF-IDF dengan algoritma *K-Means* digunakan untuk mengklaster jawaban uraian mahasiswa, yang terbukti mampu mempercepat proses penilaian ujian [4]. Widaningrum et al. [5] juga menerapkan kombinasi TF-IDF dan *K-Means clustering* untuk mengelompokkan dokumen berdasarkan kemiripan konten, sehingga mempermudah proses kategorisasi Remawati et al. [6] memanfaatkan pendekatan serupa dalam pengelompokan film trending di *YouTube*, yang memberikan gambaran mengenai preferensi penonton berdasarkan tema dan tingkat popularitas *video*.

Metode *clustering* seperti *K-Means* telah banyak dimanfaatkan untuk mengelompokkan dokumen berdasarkan kesamaan fitur yang diekstraksi, termasuk dalam analisis publikasi ilmiah. Purniawan et al. [7] menerapkan algoritma *K-Means* pada data berita daring dari situs Detik.com dan berhasil membentuk 27 klaster dari 124.509 artikel, yang menunjukkan kemampuan metode ini dalam menangani data berukuran besar. Pendekatan serupa dilakukan oleh Maulana et al. [8] dengan menggunakan metode CLARA untuk mengekstraksi topik dari dokumen berita, yang

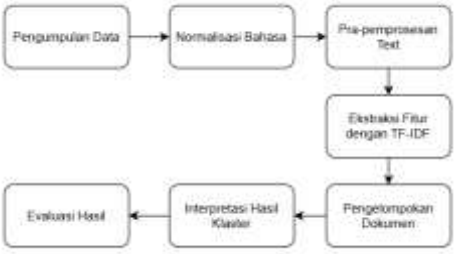
Buletin Ilmiah Informatika Teknologi
Vol 3, No 3, Mei 2025, Hal. 78 - 84
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.122
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

terbukti efektif dalam mengelompokkan dan mengidentifikasi tema utama. Dalam ranah dokumen ilmiah, Wardy et al. [9] mengombinasikan TF-IDF, PCA, dan *K-Means* untuk klasifikasi artikel, yang menghasilkan nilai akurasi dan F1-*score* yang tinggi, menandakan keandalan pendekatan tersebut dalam mengenali pola topik. Sementara itu, Simanjuntak et al. [10] menerapkan *Word Embedding* dan *K-Means* pada ratusan ribu artikel dari portal berita dan memperoleh nilai *silhouette coefficient* sebesar 0,73, yang mencerminkan kualitas pengelompokan yang baik. Berbagai studi tersebut menunjukkan bahwa kombinasi teknik TF-IDF dan *K-Means* memiliki potensi besar dalam analisis topik dominan, khususnya pada publikasi ilmiah di bidang ilmu komputer, karena kemampuannya dalam mengelompokkan dokumen berdasarkan kemiripan isi dan struktur tematik secara efektif.

Meskipun berbagai penelitian telah menerapkan kombinasi TF-IDF dan metode *clustering* untuk analisis dokumen, masih terdapat keterbatasan dalam mengidentifikasi tren topik secara dinamis dan otomatis. Penelitian yang dilakukan oleh Widaningrum et al. [5] menggunakan gabungan TF-IDF dan *K-Means Clustering* untuk menentukan kategori dokumen, namun metode tersebut belum sepenuhnya mampu menangkap perubahan topik yang terjadi seiring waktu. Berdasarkan hal tersebut, penelitian ini bertujuan untuk menganalisis topik-topik dominan dalam publikasi ilmiah bidang ilmu komputer dengan menerapkan teknik vektorisasi TF-IDF dan algoritma *clustering K-Means*. Kontribusi hasil dari penelitian ini dapat memberikan wawasan yang lebih komprehensif mengenai perkembangan tema penelitian di bidang ilmu komputer serta menjadi dasar dalam pengambilan keputusan strategis di lingkungan akademik.

2. METODOLOGI PENELITIAN

Penelitian ini mengadopsi pendekatan kuantitatif dengan menerapkan metode *text mining* untuk mengidentifikasi topik dominan dari kumpulan publikasi ilmiah di bidang ilmu komputer. Prosedur penelitian dilakukan melalui tujuh tahapan utama, mulai dari pengumpulan data, normalisasi bahasa, pra-pemrosesan teks, ekstraksi fitur dengan TF-IDF, pengelompokan dokumen, interpretasi hasil klaster, dan evaluasi hasil. Tahapan utama penelitian dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Utama Penelitian

2.1 Pengumpulan Data

Penelitian ini diawali dengan proses pengumpulan data berupa judul publikasi ilmiah di bidang ilmu komputer. Sumber data berasal dari repositori terbuka *Semantic Scholar*, dengan cakupan waktu antara tahun 2020 hingga 2025. Sebanyak 1.222 judul dokumen berhasil dikumpulkan, meskipun jumlah ini dapat disesuaikan berdasarkan kebutuhan analisis dan kapasitas pemrosesan data yang tersedia. Pendekatan ini sejalan dengan praktik dalam studi *scientometrik*, yang memanfaatkan repositori terbuka untuk keperluan analisis *bibliometrik* dan pemetaan topik penelitian [11].

2.2 Normalisasi Bahasa

Normalisasi bahasa dalam penelitian ini merujuk pada proses penerjemahan seluruh judul publikasi ilmiah yang berbahasa asing ke dalam Bahasa Indonesia. Langkah ini bertujuan untuk memastikan konsistensi linguistik dan mempermudah proses analisis teks selanjutnya. Penerjemahan dilakukan menggunakan library Python deep-translator, yang mendukung berbagai layanan terjemahan otomatis seperti Google Translate.Proses normalisasi ini penting dalam studi text mining, terutama ketika data yang digunakan berasal dari berbagai sumber dengan bahasa yang berbeda. Penelitian oleh Ardinata, Permana, dan Wijaya [12] menekankan pentingnya normalisasi teks dalam pengolahan data teks berbahasa Indonesia, khususnya dalam konteks media sosial, untuk meningkatkan akurasi analisis lebih lanjut. Mereka menggunakan metode FastText untuk identifikasi dan normalisasi teks slang pada Twitter dalam Bahasa Indonesia, yang menunjukkan peningkatan akurasi dalam analisis sentimen.

2.3 Pra-pemrosesan Teks

Setelah data terkumpul, langkah selanjutnya adalah tahap pra-pemrosesan teks untuk mempersiapkan data agar dapat dianalisis secara komputasional. Tahapan ini mencakup beberapa proses penting, antara lain:

a. *Tokenisasi*, yaitu memecah teks menjadi unit-unit kata atau token;
b. *Lowercasing*, yaitu mengubah seluruh huruf menjadi huruf kecil untuk menjaga konsistensi;

Buletin Ilmiah Informatika Teknologi
Vol 3, No 3, Mei 2025, Hal. 78 - 84
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.122
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

c. *Stopword Removal*, yaitu menghapus kata-kata umum yang tidak memiliki makna signifikan, seperti "yang", "dan", atau "adalah";
d. *Stemming*, yaitu mengembalikan kata ke bentuk dasarnya menggunakan algoritma *stemmer* Bahasa Indonesia, seperti *Sastrawi*;
e. *Filtering*, yaitu menghapus angka, tanda baca, dan karakter khusus lainnya yang tidak relevan.

Proses pra-pemrosesan ini sangat krusial untuk meningkatkan kualitas data sebelum dilakukan analisis lebih lanjut. Penelitian oleh Wardhani et al. [13] menunjukkan bahwa penerapan *stemming* dan *stopword removal* secara bersamaan mampu meningkatkan akurasi analisis sentimen hingga 93,43%. Selain itu, Santosa et al. [14] juga menekankan pentingnya tahap ini dalam meningkatkan performa klasifikasi teks secara keseluruhan.

2.4 Ekstraksi Fitur dengan TF-IDF

Setelah melalui tahap pra-pemrosesan, dokumen teks kemudian dikonversi ke dalam bentuk representasi numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). Metode ini berfungsi untuk menghitung bobot setiap kata berdasarkan dua aspek utama: frekuensi kemunculan kata dalam sebuah dokumen (*term frequency*) dan tingkat kelangkaan kata tersebut di seluruh korpus dokumen (*inverse document frequency*). Dengan demikian, kata-kata yang muncul secara signifikan dalam satu dokumen namun jarang ditemukan dalam dokumen lainnya akan memiliki bobot yang lebih tinggi, menjadikannya indikator yang kuat terhadap fitur dominan dokumen tersebut. Pendekatan ini telah terbukti efektif dalam berbagai penelitian, salah satunya oleh Widaningrum et al. [5], yang memanfaatkan TF-IDF untuk mentransformasikan teks menjadi vektor numerik sebagai dasar pengelompokan dokumen.

2.5 Pengelompokan Dokumen

Setelah dokumen direpresentasikan dalam bentuk vektor melalui proses ekstraksi fitur, tahap berikutnya adalah melakukan *clustering* menggunakan algoritma *K-Means*. Penentuan jumlah klaster yang optimal dilakukan dengan memanfaatkan dua pendekatan, yaitu metode *Elbow* dan analisis *Silhouette Score*. Metode *Elbow* dilakukan dengan menghitung nilai *Within-Cluster Sum of Squares* (WCSS) untuk berbagai nilai k, kemudian mengidentifikasi titik di mana penurunan WCSS mulai melambat secara signifikan yang dikenal sebagai "titik siku". Di sisi lain, *Silhouette Score* digunakan untuk mengevaluasi seberapa baik suatu data cocok berada dalam klasternya sendiri dibandingkan dengan klister lain, dengan rentang nilai antara -1 hingga 1. Nilai yang lebih tinggi menunjukkan kualitas pengelompokan yang lebih baik.

Setelah jumlah klaster yang paling sesuai ditentukan, algoritma *K-Means* dijalankan secara iteratif, dimulai dengan inisialisasi *centroid* secara acak dan dilanjutkan dengan pembaruan posisi *centroid* hingga mencapai konvergensi. Hasil akhir dari proses ini adalah pengelompokan dokumen ke dalam klaster-klaster yang masing-masing merepresentasikan topik dominan dalam kumpulan data. Pendekatan ini telah digunakan dalam berbagai penelitian, salah satunya oleh Muttaqin dan Defriani [15] yang berhasil mengelompokkan topik skripsi mahasiswa menggunakan algoritma *K-Means*.

2.6 Interpretasi Hasil Klaster

Setelah proses klasterisasi dokumen selesai dilakukan, tahap selanjutnya adalah menganalisis setiap klaster dengan mengekstraksi kata-kata kunci yang memiliki bobot tertinggi pada pusat klaster (*centroid*). Kata-kata tersebut kemudian dianalisis secara semantik untuk menginterpretasikan topik utama yang merepresentasikan masing-masing kelompok dokumen. Pendekatan ini selaras dengan metode yang digunakan dalam penelitian oleh Simanjuntak et al. [10], yang mengombinasikan algoritma *K-Means* dengan teknik *Word Embedding* guna meningkatkan kualitas klasterisasi serta pemahaman terhadap representasi topik dari kumpulan dokumen.

2.7 Evaluasi Hasil

Kualitas hasil klasterisasi dievaluasi menggunakan *Silhouette Score*, yang mengukur sejauh mana objek-objek dalam suatu klaster memiliki kohesi yang baik serta seberapa terpisah klaster-klaster tersebut satu sama lain. Nilai *Silhouette Score* yang mendekati angka 1 menandakan bahwa objek-objek telah dikelompokkan dengan tepat dalam klasternya, sedangkan nilai yang mendekati -1 mengindikasikan potensi kesalahan dalam pengelompokan. Pendekatan evaluasi ini juga diterapkan dalam penelitian oleh Hasan [16], yang membandingkan kinerja algoritma *K-Means* dan DBSCAN dengan menggunakan *Silhouette Score* serta *Davies-Bouldin Index* sebagai metrik penilaian.

3. HASIL DAN PEMBAHASAN

Hasil analisis topik dari dokumen ilmiah di bidang ilmu komputer dengan memanfaatkan metode TF-IDF untuk ekstraksi fitur dan algoritma *K-Means* dalam proses pengelompokan. Analisis hasil dilakukan dari dua perspektif utama, yakni jumlah dokumen dalam setiap klaster yang mencerminkan dominasi topik tertentu, serta kualitas pengelompokan yang diukur melalui evaluasi kuantitatif. Penilaian kualitas klaster menggunakan dua metrik, yaitu

Buletin Ilmiah Informatika Teknologi
Vol 3, No 3, Mei 2025, Hal. 78 - 84
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.122
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

Silhouette Score untuk menilai konsistensi internal klaster dan *Davies-Bouldin Index* untuk mengukur tingkat pemisahan antar klaster.

3.1 Hasil *Clustering* berdasarkan Jumlah Data per *Cluster*

Hasil awal yang disajikan menunjukkan sebaran jumlah dokumen atau paper dalam tiap klaster. Gambar 2 menampilkan urutan klaster berdasarkan jumlah dokumen terbanyak, sehingga dapat dilihat klaster mana yang memiliki dominasi topik lebih kuat.

Tabel 1. Ranking Berdasarkan Jumlah Data per *Cluster*

Rank	*Cluster*	Jumlah	Topik
1	4	369	keamanan, informatika, pembelajaran
2	2	184	kecerdasan, buatan, kecerdasan buatan
3	7	183	data, ilmu, ilmu data
4	0	129	pembelajaran mesin, mesin, pembelajaran
5	3	89	peretasan, etis, peretasan etis
6	5	75	deteksi intrusi, intrusi, deteksi
7	1	71	bahasa, bahasa alami, alami
8	8	63	pengembangan, lunak, perangkat lunak
9	6	59	visi, visi komputer, komputer

Dari Tabel 1 tersebut, dapat dilihat bahwa:

a. *Cluster* 4 memiliki jumlah dokumen terbanyak, yaitu 369 dokumen, dengan topik dominan keamanan, informatika, pembelajaran. Hal ini menunjukkan bahwa tema keamanan informasi dan teknologi pembelajaran menjadi topik yang paling sering diteliti dalam kumpulan data ini.
b. Disusul oleh *Cluster* 2 yang berisi 184 dokumen, dengan topik utama kecerdasan buatan. Ini menunjukkan tingginya perhatian terhadap perkembangan teknologi artificial intelligence.
c. *Cluster* 3 memiliki 183 dokumen dan mencakup topik data dan ilmu data, yang menunjukkan bahwa analisis dan pemrosesan data merupakan bidang aktif dalam penelitian ilmu komputer.
d. *Cluster* 0 memuat 129 dokumen dengan tema pembelajaran mesin dan mesin, yang menegaskan bahwa machine learning merupakan subbidang yang juga sangat berkembang.
e. Klaster-klaster lainnya seperti *Cluster* 5 (89 dokumen) dan *Cluster* 7 (71 dokumen) membahas topik seperti peretasan etis dan bahasa alami, menunjukkan adanya minat pada isu keamanan dan pemrosesan bahasa.
f. Sementara itu, *Cluster* 6 menjadi klaster dengan jumlah dokumen paling sedikit, yaitu 59 dokumen, dengan topik terkait visi komputer dan komputer, mengindikasikan bahwa tema ini masih relatif minor dalam kumpulan data yang dianalisis.

Distribusi ini menunjukkan bahwa topik-topik populer seperti keamanan siber, pembelajaran mesin, dan kecerdasan buatan memiliki representasi dokumen yang lebih tinggi, sementara beberapa topik spesifik seperti visi komputer dan pengembangan perangkat lunak memiliki representasi yang lebih rendah.

3.2 Evaluasi Kualitas *Cluster* Berdasarkan *Silhouette Score*

Untuk menilai seberapa baik hasil *clustering*, digunakan *metrik Silhouette Score*. *Metrik* ini mengukur sejauh mana sebuah dokumen cocok dengan klaster tempat ia berada dibandingkan dengan klaster lainnya. Nilai *Silhouette* berkisar dari -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan kohesi yang lebih baik di dalam klaster dan separasi yang lebih jelas terhadap klaster lainnya.

Tabel 2. Ranking Kualitas *Cluster* berdasarkan *Silhouette Score*

Rank	*Cluster*	*Silhouette*	Topik
1	1	0.183500	bahasa, bahasa alami, alami
2	6	0.115800	visi, visi komputer, komputer
3	5	0.112200	deteksi intrusi, intrusi, deteksi
4	3	0.106800	peretasan, etis, peretasan etis
5	8	0.079000	pengembangan, lunak, perangkat lunak
6	2	0.07300	kecerdasan, buatan, kecerdasan buatan
7	0	0.070200	pembelajaran mesin, mesin, pembelajaran
8	7	0.046200	data, ilmu, ilmu data
9	4	-.0006200	keamanan, informatika, pembelajaran

Berdasarkan Tabel 2, klaster yang memiliki kualitas paling baik adalah:

a. *Cluster* 1 dengan nilai *Silhouette Score* sebesar 0.1835, yang memuat topik bahasa, bahasa alami, alami. Nilai ini menunjukkan bahwa klaster tersebut memiliki kohesi internal yang tinggi dan terpisah cukup baik dari klaster lainnya.

Buletin Ilmiah Informatika Teknologi
Vol 3, No 3, Mei 2025, Hal. 78 - 84
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.122
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

b. Disusul oleh *Cluster* 6 (0.1158) yang berkaitan dengan topik visi komputer, serta *Cluster* 5 (0.1122) yang mencakup deteksi intrusi. Ketiganya menunjukkan kualitas klasterisasi yang relatif baik.
c. Klaster lainnya seperti *Cluster* 3 (0.1068) dan *Cluster* 8 (0.0790) juga masih berada pada kisaran positif, menandakan bahwa meskipun tidak sekuat klaster terbaik, struktur klasternya masih cukup valid.
d. Sementara itu, *Cluster* 4, meskipun memiliki jumlah dokumen terbanyak seperti ditunjukkan pada Gambar 2, justru memiliki *Silhouette Score negatif* (-0.0062). Ini mengindikasikan adanya tumpang tindih yang tinggi dengan klaster lain dan kemungkinan besar kurang representatif secara tematik.

Temuan ini menegaskan bahwa jumlah dokumen dalam suatu klaster tidak selalu sejalan dengan kualitas klaster tersebut. Klaster besar bisa jadi mengandung variasi topik yang terlalu luas sehingga sulit dibedakan secara jelas dari klaster lain. Oleh karena itu, penilaian berbasis metrik kuantitatif seperti *Silhouette Score* penting untuk mengevaluasi kualitas segmentasi topik secara objektif.

3.3 Evaluasi Hasil

Evaluasi secara keseluruhan terhadap performa *clustering* dilakukan dengan dua metrik:

a. *Silhouette Score Global*: mengukur konsistensi data dalam setiap klaster dengan membandingkan jarak antar dokumen dalam klaster yang sama dan dokumen di klaster lain. Nilai ini berkisar antara -1 sampai 1, di mana nilai yang lebih tinggi menunjukkan klaster yang lebih jelas dan terpisah.
b. *Davies-Bouldin Index* (DBI): mengukur seberapa baik pemisahan antar klaster dengan membandingkan jarak antar pusat klaster dan ukuran klaster itu sendiri. Nilai DBI yang lebih rendah menunjukkan kualitas klaster yang lebih baik dengan pemisahan yang jelas.

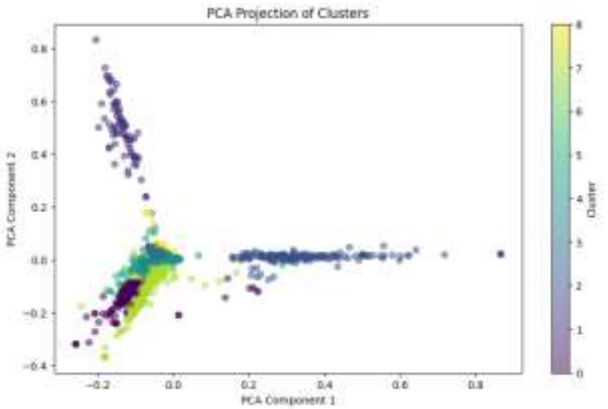
Tabel 3. Evaluasi Kualitas *Clustering Global*

Metode	Nilai
Silhouette Score	0.058500
Davies-Bouldin index	4.387000

Nilai yang diperoleh dari evaluasi dapat dilihat pada Tabel 3. yaitu :

a. *Silhouette Score*: 0.0585
b. *Davies-Bouldin Index*: 4.387

Nilai *silhouette* yang relatif rendah ini menandakan bahwa sebagian besar dokumen berada di area perbatasan antara dua klaster, sehingga pengelompokan tidak sangat jelas dan masih terdapat banyak dokumen yang berada di antara klaster. Sedangkan nilai DBI yang cukup tinggi mengindikasikan adanya tumpang tindih antar klaster yang cukup signifikan. Kondisi ini kemungkinan besar dipengaruhi oleh factor fitur yang digunakan, yaitu TF-IDF, bersifat *sparse* dan hanya merepresentasikan frekuensi kata tanpa memperhitungkan konteks semantik, sehingga kemiripan topik yang lebih halus sulit dideteksi.



Gambar 2. Visualisasi PCA

Visualisasi hasil klasterisasi menggunakan algoritma K-Means dapat dilihat pada Gambar 2 yang menampilkan distribusi dokumen berdasarkan klaster dalam ruang dua dimensi hasil reduksi PCA. Setiap warna mewakili klaster berbeda. Dari visualisasi ini terlihat beberapa klaster memiliki pemisahan yang cukup jelas, menandakan dokumen dalam klaster tersebut memiliki kesamaan topik yang kuat. Namun, terdapat juga beberapa klaster yang saling berdekatan dan tumpang tindih, yang mengindikasikan kemiripan topik di antara kelompok dokumen tersebut.

Selain evaluasi hasil *clustering*, perlu dicatat bahwa proses normalisasi bahasa yang dilakukan pada data teks memakan waktu yang cukup lama. Untuk jumlah data sebanyak 1222 judul dokumen, proses normalisasi Bahasa membutuhkan waktu sekitar 24 menit. Hal ini dikarenakan *library* yang digunakan untuk melakukan translate teks adalah *deep-translator* yang sangat bergantung pada kecepatan koneksi internet dan batasan dari *Google* Translate API publik.

Buletin Ilmiah Informatika Teknologi
Vol 3, No 3, Mei 2025, Hal. 78 - 84
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.122
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

3.4 Pembahasan

Hasil analisis menunjukkan berbagai temuan penting terkait performa dan kualitas klaster yang terbentuk dari dokumen-dokumen ilmiah bidang ilmu komputer:

a. Distribusi dokumen tidak selalu mencerminkan kualitas klaster. Meskipun *Cluster* 4 merupakan klaster dengan jumlah dokumen terbanyak (369 dokumen) dengan topik dominan keamanan, informatika, dan pembelajaran, kualitasnya justru paling rendah berdasarkan *Silhouette Score* (-0.0062). Ini menandakan bahwa topik-topik dalam klaster tersebut terlalu umum dan tumpang tindih dengan klaster lain.
b. Klaster dengan topik spesifik cenderung memiliki kualitas lebih baik. *Cluster* 1, dengan topik bahasa alami, dan *Cluster* 6, bertopik visi komputer, memperoleh nilai *Silhouette* tertinggi (masing-masing 0.1835 dan 0.1158). Hal ini menunjukkan bahwa dokumen dalam klaster ini lebih konsisten dan homogen.
c. Evaluasi global mengindikasikan pemisahan klaster yang kurang optimal. Nilai *Silhouette Score* global sebesar 0.0585 tergolong rendah, menandakan bahwa sebagian besar dokumen berada di dekat batas antara klaster. Nilai *Davies-Bouldin Index* yang tinggi (4.387) juga menunjukkan bahwa klaster-klaster yang terbentuk belum memiliki separasi yang baik dan saling tumpang tindih.
d. Visualisasi PCA memperkuat hasil evaluasi numerik. Dari proyeksi dua dimensi menggunakan PCA, terlihat bahwa sebagian klaster memang membentuk gugus yang cukup terpisah, namun terdapat pula area di mana klaster-klaster saling berdekatan dan menyatu. Hal ini mengindikasikan bahwa representasi berbasis TF-IDF belum cukup menangkap dimensi semantik secara menyeluruh.

Secara keseluruhan, hasil ini menunjukkan bahwa meskipun beberapa topik telah terklasterisasi dengan baik, performa *clustering* secara global masih perlu ditingkatkan. Hal ini membuka peluang untuk penggunaan teknik representasi fitur yang lebih kaya konteks, seperti *word embeddings* atau *contextual embeddings*, serta pendekatan *clustering* yang lebih adaptif.

4. KESIMPULAN

Penelitian ini mengungkap bahwa penerapan metode TF-IDF dan *K-Means* berhasil mengelompokkan dokumen ilmiah bidang ilmu komputer ke dalam sejumlah klaster topik, dengan dominasi pada tema-tema seperti keamanan siber, kecerdasan buatan, dan pembelajaran mesin. Namun demikian, hasil evaluasi menunjukkan bahwa distribusi jumlah dokumen dalam klaster tidak selalu mencerminkan kualitas pengelompokan. Klaster dengan topik yang lebih spesifik, seperti bahasa alami (*Cluster* 1) dan visi komputer (*Cluster* 6), cenderung memiliki nilai Silhouette Score yang lebih tinggi, menunjukkan kohesi internal yang baik. Sebaliknya, klaster dengan cakupan topik yang luas seperti *Cluster* 4, meskipun berisi dokumen terbanyak, justru memiliki kualitas pengelompokan yang rendah dan tumpang tindih dengan klaster lain. Evaluasi global melalui *Silhouette Score* sebesar 0.0585 dan *Davies-Bouldin Index* sebesar 4.387 mengindikasikan bahwa pemisahan antar klaster masih kurang optimal. Visualisasi menggunakan PCA turut memperkuat temuan tersebut dengan menunjukkan adanya klaster yang tumpang tindih secara visual. Temuan ini menegaskan bahwa penggunaan TF-IDF sebagai representasi fitur memiliki keterbatasan dalam menangkap makna semantik yang lebih dalam antar dokumen. Oleh karena itu, disarankan untuk mengeksplorasi pendekatan representasi fitur yang lebih kontekstual, seperti *word embeddings* atau *contextual embeddings*, serta algoritma *clustering* yang lebih adaptif untuk meningkatkan akurasi dan kualitas segmentasi topik.

REFERENCES

[1] N. W. Utami and I. G. J. Eka Putra, "Text Minig Clustering Untuk Pengelompokan Topik Dokumen Penelitian Menggunakan Algoritma K-Means Dengan Cosine Similarity," *J. Inform. Teknol. dan Sains*, vol. 4, no. 3, pp. 255–259, 2022, doi: 10.51401/jinteks.v4i3.1907.

[2] M. A. Haq, W. Purnomo, and N. Y. Setiawan, "Analisis Clustering Topik Survey menggunakan Algoritme K-Means (Studi Kasus: Kudata)," *... Teknol. Inf. dan Ilmu ...*, vol. 7, no. 7, pp. 3498–3506, 2023, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/13147%0Ahttps://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/13147/5928

[3] D. A. C. Rachman, R. Goejantoro, and F. D. T. Amijaya, "Implementasi Text Mining Pengelompokkan Dokumen Skripsi Menggunakan Metode K-Means Clustering," *Eksponensial*, vol. 11, no. 2, p. 167, 2020, doi: 10.30872/eksponensial.v11i2.660.

[4] I. A. Mashudi, S. N. Arief, D. S. E.I., T. Fatmawati, M. Hani'ah, and I. T. Alfarid, "Klasterisasi Jawaban Uraian Mahasiswa Menggunakan TF-IDF dan K-Means untuk Membantu Koreksi Ujian," *J. Media Inform. Budidarma*, vol. 7, no. 4, p. 2159, 2023, doi: 10.30865/mib.v7i4.6688.

[5] I. Widaningrum, D. Mustikasari, R. Arifin, S. L. Tsaqila, and D. Fatmawati, "Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) dan K-Means Clustering Untuk Menentukan Kategori Dokumen," *Pros. Semin. Nas. Sist. Inf. dan Teknol.*, pp. 145–149, 2022.

[6] K. Clastering, D. Remawati, H. Wijayanto, Y. Retno, W. Utami, and B. D. Raharja, "Pengelompokkan Film Trending di Youtube Menggunakan TF-IDF dan," vol. 4, pp. 65–74, 2025.

[7] I. M. A. Purniawan, G. M. A. Sasmita, and I. P. A. E. Pratama, "Clustering Berita Menggunakan Algoritma Tf-Idf Dan K-Means Dengan Memanfaatkan Sumber Data Crawling Pada Situs Detik.Com," *JITTER- J. Ilm. Teknol. dan Komput.*, vol. 3, no. 1, pp. 821–830, 2022.

Buletin Ilmiah Informatika Teknologi
Vol 3, No 3, Mei 2025, Hal. 78 - 84
ISSN 2962-0945 (media online)
DOI 10.58369/biit.v2i3.122
https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT

[8] R. Maulana and S. Adinugroho, "Ekstraksi Topik Dokumen Berita Menggunakan Term-Cluster Weighting dan Clustering Large Application (CLARA)," vol. 3, no. 11, pp. 10623–10629, 2019, [Online]. Available: http://j-ptiik.ub.ac.id

[9] D. K. Wardy, I. K. G. D. Putra, and N. K. D. Rusjayanthi, "Clustering Artikel pada Portal Berita Online," *JITTER- J. Ilm. Teknol. dan Komput.*, vol. 3, no. 1, pp. 3–11, 2022.

[10] H. T. A. Simanjuntak, P. E. P. Silaban, J. K. S. Manurung, and V. H. Sormin, "Klasterisasi Berita Bahasa Indonesia Dengan Menggunakan K-Means Dan Word Embedding," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 3, pp. 641–652, 2023, doi: 10.25126/jtiik.20231026468.

[11] N. R. Rosiyan, "Pemetaan Sistematik Publikasi Tren Penelitian Pustakawan Data Menggunakan ScientoPy," *Media Pustak.*, vol. 30, no. 3, pp. 235–244, 2023, doi: 10.37014/medpus.v30i3.4954.

[12] Pande sindu, Agus Aan Jiwa Permana, and I Nyoman Saputra Wahyu Wijaya, "Identifikasi Dan Normalisasi Teks Slang Dengan Fasttext Pada Twitter Dalam Bahasa Indonesia," *J. Pendidik. Teknol. dan Kejuru.*, vol. 21, no. 1, pp. 33–44, 2024, doi: 10.23887/jptkundiksha.v21i1.66381.

[13] S. Analisis, A. Satusehat, D. Wardhani, R. Astuti, and D. D. Saputra, "Optimasi Feature Selection Text Mining: Stemming dan Stopword," *Innov. J. Soc. Sci. Res.*, vol. 4, pp. 7537–7548, 2024.

[14] A. Santosa, I. Purnamasari, and Mayasari Rini, "Pengaruh Stopword Removal dan StemmingTerhadap Performa Klasifikasi Teks KomentarKebijakan New Normal Menggunakan AlgoritmaLSTM," *J. Sains Komput. Inform.*, vol. 6, pp. 81–93, 2022.

[15] M. R. Muttaqin and M. Defriani, "Algoritma K-Means untuk Pengelompokan Topik Skripsi Mahasiswa," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 121–129, 2020, doi: 10.33096/ilkom.v12i2.542.121-129.

[16] K. Dbscan and Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index pada Hasil Cluster," vol. 06, no. 01, pp. 60–74, 2024.