Pengelompokkan Abstrak Jurnal Ilmiah Menggunakan Term Frequency– Inverse Document Frequency dan K-Means

Nadia Wati Aprianti[1], **Herman Yuliansyah**[2*], **Muhammad Kunta Biddinika**[3]
[1,2,3] Program Studi Magister Informatika, Fakultas Teknologi Industri, Universitas Ahmad Dahlan
[*]Email: 2408048016@webmail.uad.ac.id

DOI: https://doi.org/10.31603/komtika.v10i1.15516

ABSTRACT

The rapid growth of scientific publications in the field of informatics has created challenges in efficiently clustering and analyzing article abstracts. Document clustering based on topic similarity is essential to support more structured information retrieval, management, and analysis. This study aims to propose a machine learning–based clustering model for grouping informatics journal article abstracts according to topic similarity. The dataset consists of 1,200 English language journal abstracts obtained from the SINTA portal for the 2023–2024 period. The text preprocessing stages include case folding, tokenization, stop word removal, and stemming. Document representation is constructed using the Term Frequency–Inverse Document Frequency (TF-IDF) method, followed by clustering using the K-Means algorithm. The optimal number of clusters is determined through evaluation using the Silhouette Score and Davies–Bouldin Index. The results show that the optimal clustering is achieved with 19 clusters, yielding a Silhouette Score of 0.0708, which indicates relatively weak cluster separation an outcome commonly observed in high dimensional text data and a Davies–Bouldin Index of 3.2760, indicating a reasonably good level of cluster compactness and separation. These findings suggest that TF-IDF based feature representation provides adequate capability in supporting the clustering of informatics journal article abstracts based on topic similarity.

Keywords: *Abstract clustering, K-means, Latent dirichlet allocation, Text mining, TF-IDF.*

ABSTRAK

Pertumbuhan publikasi ilmiah di bidang informatika menimbulkan tantangan dalam proses pengelompokan dan analisis abstrak artikel secara efisien. Pengelompokan dokumen berdasarkan kemiripan topik diperlukan untuk mendukung pencarian, pengelolaan, dan analisis informasi secara lebih terstruktur. Penelitian ini bertujuan untuk mengusulkan model clustering berbasis machine learning dalam pengelompokan abstrak artikel jurnal informatika berdasarkan kemiripan topik. Dataset penelitian terdiri atas 1200 abstrak artikel jurnal berbahasa Inggris yang diperoleh dari portal SINTA periode 2023–2024. Tahapan prapemrosesan teks meliputi case folding, tokenisasi, stopword removal, dan stemming. Representasi dokumen dibangun menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF), kemudian dilakukan pengelompokan menggunakan algoritma K-Means. Penentuan jumlah klaster optimal dilakukan melalui evaluasi menggunakan Silhouette Score dan Davies–Bouldin Index. Hasil penelitian menunjukkan bahwa jumlah klaster optimal diperoleh pada 19 klaster dengan nilai Silhouette Score sebesar 0,0708, yang mengindikasikan pemisahan klaster yang relatif lemah suatu kondisi yang umum terjadi pada data teks berdimensi tinggi serta nilai Davies–Bouldin Index sebesar 3,2760 yang menunjukkan tingkat kekompakan dan pemisahan klaster yang cukup baik. Dengan demikian, representasi fitur berbasis TF-IDF menunjukkan kemampuan yang memadai dalam mendukung proses pengelompokan abstrak artikel jurnal informatika berdasarkan kemiripan topik.

Kata kunci: Klustering abstrak, K-means, Latent dirichlet allocation, Text mining, TF-IDF.

PENDAHULUAN

Pertumbuhan dokumen teks digital, khususnya artikel ilmiah, menyebabkan proses analisis manual menjadi kurang efisien karena besarnya volume data dan sifat data yang tidak

terstruktur. Kondisi tersebut mendorong perlunya pendekatan otomatis untuk mengekstraksi informasi penting dari kumpulan dokumen dalam jumlah besar [1]. Salah satu metode yang kerap diterapkan dalam analisis teks adalah *text mining*, yakni pendekatan yang bertujuan mengekstraksi informasi dan pola tersembunyi dari kumpulan data teks yang tidak terstruktur. Sebelum data teks dapat dianalisis lebih lanjut, diperlukan serangkaian tahapan prapemrosesan (*preprocessing*) yang mencakup *case folding*, tokenisasi, *stopword removal*, dan *stemming*. Tahapan ini berperan krusial dalam membersihkan data dari *noise* serta memperkuat kualitas representasi teks. Selanjutnya, dokumen hasil prapemrosesan perlu dikonversi ke dalam format numerik agar dapat diolah secara komputasional oleh algoritma pembelajaran mesin.[2], [3].

Representasi fitur teks dilakukan menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF), yang bekerja dengan menghitung bobot setiap kata berdasarkan frekuensi kemunculannya pada suatu dokumen dibandingkan dengan distribusinya di seluruh korpus. Pendekatan ini efektif dalam membedakan kata-kata yang bersifat informatif dari kata-kata yang terlalu umum, sehingga mampu mencerminkan tingkat relevansi dan kepentingan suatu istilah dalam konteks dokumen secara lebih akurat. [4].

Dalam pengolahan dokumen teks berskala besar, *clustering* merupakan salah satu teknik analisis yang banyak diterapkan, yaitu suatu proses untuk mengelompokkan dokumen-dokumen yang memiliki kemiripan karakteristik ke dalam klaster yang sama secara otomatis. Penerapan metode ini memberikan kemudahan dalam pengelolaan, analisis, maupun pencarian informasi karena dokumen dengan topik yang saling berkaitan dapat terhimpun dalam satu kelompok tanpa memerlukan pelabelan manual. Di antara berbagai algoritma *clustering* yang tersedia, K-Means menjadi pilihan yang paling banyak diterapkan mengingat efisiensi komputasinya yang tinggi, kemudahan dalam implementasi, serta kemampuannya dalam memproses dataset berukuran besar dengan baik [5], [6]. Namun demikian, perlu diperhatikan bahwa kualitas pengelompokan yang dihasilkan sangat bergantung pada kualitas representasi fitur yang digunakan, terutama mengingat data teks secara inheren memiliki dimensi yang sangat tinggi dan bersifat *sparse*. Kondisi tersebut menjadi tantangan tersendiri dalam proses *clustering* dokumen teks. Untuk mengatasi hal ini, kombinasi antara TF-IDF sebagai metode representasi fitur dan K-Means sebagai algoritma pengelompokan dipandang sebagai pendekatan yang tepat dan efektif dalam menghasilkan klaster dokumen berdasarkan kemiripan topik penelitian.

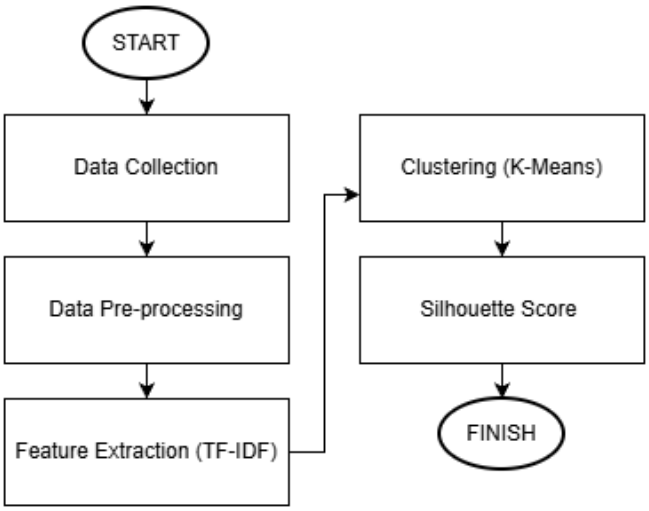
Berbagai penelitian sebelumnya telah menerapkan TF-IDF dan K-Means dalam pengelompokan dokumen teks pada berbagai domain, seperti berita, media sosial, dan dokumen akademik[7], [8]. Namun, penelitian yang secara khusus berfokus pada clustering abstrak artikel jurnal informatika dari portal SINTA masih relatif terbatas, terutama pada data publikasi terbaru periode 2023–2024. Selain itu, sebagian penelitian terdahulu lebih menitikberatkan pada proses klasifikasi dokumen dibandingkan eksplorasi struktur topik menggunakan pendekatan clustering[9]. Oleh karena itu, penelitian ini berupaya mengisi celah tersebut dengan menerapkan kombinasi TF-IDF dan K-Means untuk mengelompokkan abstrak artikel jurnal informatika berdasarkan kemiripan topik penelitian.

Sumber data penelitian ini berasal dari portal *Science and Technology Index* (SINTA), yang diperoleh melalui proses pengumpulan abstrak artikel jurnal di bidang informatika yang diterbitkan pada rentang waktu 2023 hingga 2024. Data yang terkumpul selanjutnya diproses

melalui serangkaian tahapan, meliputi prapemrosesan teks, pembobotan fitur menggunakan TF-IDF, pengelompokan dokumen menggunakan algoritma K-Means, serta pengukuran kualitas klaster menggunakan *Silhouette Score* dan *Davies–Bouldin Index* (DBI). Melalui serangkaian tahapan tersebut, penelitian ini berupaya mengusulkan model pengelompokan abstrak artikel jurnal informatika yang berbasis TF-IDF dan K-Means, dengan harapan dapat mendukung organisasi dokumen secara lebih sistematis, terstruktur, serta mampu merepresentasikan kecenderungan topik penelitian yang terdapat pada dataset yang dianalisis.

METODE

Metode penelitian ini diawali dengan tahap pengumpulan data berupa dokumen teks, Dataset penelitian diperoleh melalui proses *crawling* menggunakan kata kunci *"Informatics and computers"* pada abstrak artikel jurnal dari portal SINTA periode 2023–2024, dengan total keseluruhan dataset sebanyak 1.200 abstrak artikel ilmiah.yang kemudian diproses melalui tahapan *pre-processing* untuk meningkatkan kualitas data, meliputi normalisasi teks, tokenisasi, penghapusan *stopwords*, serta stemming atau lemmatisasi. Tahap berikutnya adalah pembentukan representasi fitur teks dengan memanfaatkan metode TF-IDF, yang bekerja dengan menetapkan nilai bobot pada setiap kata berdasarkan dua aspek sekaligus, yaitu seberapa sering kata tersebut muncul dalam sebuah dokumen dan seberapa jarang kata tersebut dijumpai di seluruh korpus. Dengan mepertimbangkan kedua aspek tersebut secara bersamaan, metode ini mampu menghasilkan representasi numerik yang mencerminkan tingkat signifikansi dan kekhasan setiap kata dalam merepresentasikan isi dokumen secara statistis [10], [11]. Vektor numerik hasil representasi fitur selanjutnya diproses menggunakan algoritma K-Means guna mengidentifikasi kelompok-kelompok dokumen yang memiliki kedekatan karakteristik dalam ruang fitur multidimensi. Proses pengelompokan ini dilakukan dengan cara mengukur kemiripan antar dokumen berdasarkan posisi vektornya, sehingga dokumen-dokumen yang memiliki profil fitur serupa dapat terhimpun ke dalam klaster yang sama [12], [13], [14]. Proses pengelompokan yang diterapkan dalam penelitian ini tergolong dalam paradigma *unsupervised learning*, di mana algoritma bekerja secara mandiri tanpa memerlukan informasi label kelas sebagai acuan, melainkan berupaya mengungkap pola dan struktur tersembunyi yang secara alami terdapat dalam data. Guna menilai sejauh mana kualitas klaster yang terbentuk, digunakan metrik *Silhouette Score* sebagai instrumen evaluasi. Metrik ini bekerja dengan mengkuantifikasi dua dimensi kualitas klaster secara bersamaan, yakni tingkat kerapatan data dalam klaster yang sama (*intra-cluster cohesion*) serta tingkat keterpisahan antara satu klaster dengan klaster lainnya (*inter-cluster separation*), sehingga memberikan gambaran yang komprehensif mengenai baik tidaknya hasil pengelompokan yang dihasilkan [6], [15], [16]. Tahap akhir penelitian dilakukan dengan menginterpretasikan hasil clustering guna memperoleh pemahaman terhadap pola dan distribusi data. Ditunjukkan pada Gambar 1.

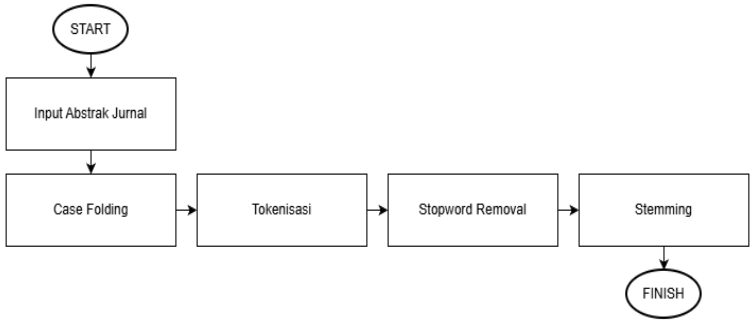


Gambar 1. Desain Penelitian

Pengumpulan data dalam penelitian ini bertujuan untuk menghimpun dataset berupa abstrak artikel ilmiah di bidang informatika yang akan dijadikan bahan analisis. Sebagai sumber data utama, digunakan portal *Science and Technology Index* (SINTA) yang merupakan basis data publikasi jurnal ilmiah berstandar akreditasi nasional di Indonesia. Untuk memperoleh data dalam jumlah besar secara sistematis, proses pengambilan data dilakukan secara otomatis dengan menerapkan teknik *web crawling* berbantuan kata kunci yang disesuaikan dengan ruang lingkup bidang informatika, guna memastikan kesesuaian data yang terkumpul dengan fokus topik penelitian. Penerapan teknik ini memungkinkan proses akuisisi data berlangsung secara lebih efisien dan terorganisir dibandingkan pengumpulan secara manual. Setelah tahap *crawling* selesai dilaksanakan, data yang berhasil dihimpun kemudian melewati proses penyaringan (*filtering*) dan seleksi untuk menjamin kelayakan serta kualitas dataset yang akan digunakan dalam analisis selanjutnya [17], [18]. Pada tahap ini, abstrak yang tidak relevan, duplikat, kosong, atau tidak sesuai dengan ruang lingkup penelitian dihapus dari dataset. Selain itu, hanya artikel jurnal yang diterbitkan pada periode tahun 2023 hingga 2024 yang digunakan dalam penelitian agar data yang dianalisis tetap relevan dan mutakhir. Hasil dari proses pengumpulan dan seleksi data menghasilkan sebanyak 1.200 abstrak artikel ilmiah berbahasa Inggris, yang terdiri atas 600 abstrak dari tahun 2023 dan 600 abstrak dari tahun 2024. Dataset tersebut selanjutnya digunakan sebagai data utama dalam tahapan preprocessing teks, ekstraksi fitur menggunakan metode TF-IDF, serta proses clustering dokumen menggunakan algoritma K-Means, Hasilnya bisa di lihat pada tabel.

Tahap *text preprocessing* dilakukan untuk mempersiapkan data abstrak artikel sebelum proses ekstraksi fitur dan clustering dilakukan. Tahapan ini bertujuan untuk membersihkan data teks, mengurangi noise, menormalkan bentuk kata, serta meningkatkan konsistensi representasi dokumen sehingga kualitas hasil clustering dapat menjadi lebih optimal. Abstrak artikel dalam kondisi mentah (*raw text*) pada umumnya masih mengandung berbagai bentuk ketidakkonsistenan, seperti variasi penggunaan huruf kapital, keberadaan tanda baca, angka, maupun simbol-simbol tertentu, serta kemunculan kata-kata yang bersifat umum dan tidak memberikan kontribusi signifikan terhadap proses analisis makna dokumen

[19], [20], [21]. Dengan demikian, tahapan prapemrosesan (*preprocessing*) memegang peranan yang sangat krusial dalam penelitian berbasis *text mining*, mengingat kualitas fitur yang dihasilkan pada tahap representasi sangat ditentukan oleh seberapa baik data teks dibersihkan dan distandarisasi pada tahap ini. Alur tahapan *preprocessing* yang diterapkan dalam penelitian ini diilustrasikan pada Gambar 2.



Gambar 2. *Text preprocessing*

Rangkaian tahapan prapemrosesan yang diterapkan dalam penelitian ini mencakup empat proses utama, yakni *case folding*, tokenisasi, *stopword removal*, dan *stemming*. Pada tahap *case folding*, seluruh karakter teks dikonversi menjadi huruf kecil sekaligus dilakukan penghapusan karakter-karakter yang tidak relevan dengan analisis. Proses tokenisasi kemudian diterapkan untuk memisahkan teks menjadi satuan-satuan kata atau token yang lebih kecil sehingga memudahkan pemrosesan lebih lanjut. Berikutnya, *stopword removal* dilaksanakan dengan cara menyaring dan membuang kata-kata bermakna umum yang dinilai tidak berkontribusi dalam merepresentasikan topik dokumen. Terakhir, proses *stemming* diterapkan guna mereduksi setiap kata ke bentuk dasarnya, sehingga berbagai variasi morfologis dari kata yang memiliki akar makna serupa dapat dilebur menjadi satu representasi yang seragam[10]. Keluaran akhir dari seluruh rangkaian tahapan prapemrosesan adalah kumpulan teks abstrak yang telah tersanitasi dan terstandarisasi secara konsisten, yang selanjutnya menjadi masukan pada tahap pembentukan representasi fitur menggunakan TF-IDF serta proses pengelompokan dokumen melalui algoritma K-Means.

Inverse Document Frequency (IDF) berfungsi untuk mengukur tingkat keunikan dan kekhususan suatu istilah dalam keseluruhan korpus dokumen. Semakin besar nilai TF(t) Semakin tinggi frekuensi kemunculan suatu istilah di berbagai dokumen dalam korpus, maka istilah tersebut dianggap semakin bersifat umum, yang pada akhirnya akan berdampak pada semakin rendahnya nilai TF-IDF yang diperoleh untuk istilah tersebut. [22], Dirumuskan pada Persamaan (1).

$$TF(t,d) = \frac{N(t,d)}{\mathrm{T}} \tag{1}$$

Inverse Document Frequency (IDF) menghitung derajat signifikansi sebuah istilah terhadap keseluruhan koleksi dokumen yang ada. Nilai *Document Frequency* (DF) diperoleh melalui perhitungan logaritma dari perbandingan antara total jumlah dokumen E dan jumlah dokumen yang mengandung istilah tersebut (N(t)) [22], Dirumuskan pada Persamaan (2).

$$IDF\ (t) = \ \mathrm{Log(E)/(N(t))} \tag{2}$$

Term Frequency (TF) dan Inverse Document Frequency (IDF). Formula ini dirancang untuk menghasilkan nilai bobot akhir sebuah istilah dalam suatu dokumen dengan memperhitungkan dua komponen secara simultan, yaitu seberapa sering istilah tersebut hadir dalam dokumen tertentu dan seberapa eksklusif kemunculannya di seluruh korpus. Apabila sebuah istilah memiliki frekuensi kemunculan yang tinggi dalam satu dokumen tertentu namun relatif jarang ditemukan pada dokumen-dokumen lainnya, maka nilai TF-IDF yang diperoleh akan cenderung tinggi, yang mengindikasikan bahwa istilah tersebut memiliki peran penting dalam mencerminkan kandungan informasi dokumen. Sebaliknya, istilah yang bersifat umum dan tersebar luas di hampir seluruh dokumen dalam korpus akan menghasilkan nilai TF-IDF yang rendah, karena dianggap kurang mampu membedakan satu dokumen dengan dokumen lainnya [22], Dirumuskan pada Persamaan (3).

$$TF - IDF = \ \mathrm{TF} * \mathrm{IDF} \tag{3}$$

Tahapan evaluasi dalam penelitian ini dilaksanakan guna menilai sejauh mana kualitas klaster yang terbentuk dari proses pengelompokan menggunakan algoritma K-Means. Dua metrik evaluasi yang digunakan secara bersamaan adalah *Silhouette Coefficient* dan *Davies–Bouldin Index* (DBI). *Silhouette Coefficient* dimanfaatkan sebagai acuan dalam menentukan jumlah klaster yang paling optimal, dengan cara mengkuantifikasi dua aspek kualitas klaster secara bersamaan, yaitu tingkat kerapatan data dalam satu klaster (*intra-cluster cohesion*) dan tingkat keterpisahan antara klaster yang berbeda (*inter-cluster separation*). Nilai *Silhouette* yang semakin mendekati angka satu mengindikasikan bahwa struktur klaster yang terbentuk semakin baik, dengan batas antar klaster yang semakin tegas dan performa pengelompokan yang semakin optimal. Secara matematis [15], [23], Silhouette Coefficient dirumuskan pada Persamaan (4).

$$S = \frac{b-a}{\max(a,b)} \tag{4}$$

Selain memanfaatkan *Silhouette Coefficient*, penelitian ini turut mengintegrasikan *Davies–Bouldin Index* (DBI) sebagai instrumen tambahan dalam mengukur kualitas hasil pengelompokan. Berbeda dengan *Silhouette Coefficient*, DBI bekerja dengan prinsip bahwa semakin kecil nilainya, semakin baik kualitas klaster yang terbentuk, karena nilai DBI yang rendah mencerminkan kondisi di mana jarak antar pusat klaster relatif jauh sementara sebaran data di dalam masing-masing klaster tetap terjaga dalam rentang yang sempit. Secara teknis, perhitungan DBI dilakukan dengan mengkuantifikasi rasio antara rata-rata jarak setiap titik data terhadap centroid klasternya masing-masing dengan jarak yang memisahkan antar centroid klaster yang berdekatan, sebagaimana dirumuskan pada Persamaan [6], sebagaimana ditunjukkan pada Persamaan (5).

$$DBI = \frac{1}{K} + \sum_{i=1}^{K} maxj \neq i \left(\frac{\sigma i + \sigma j}{dij}\right) \tag{5}$$

Algoritma K-Means diterapkan untuk membagi keseluruhan data ke dalam sejumlah k klaster

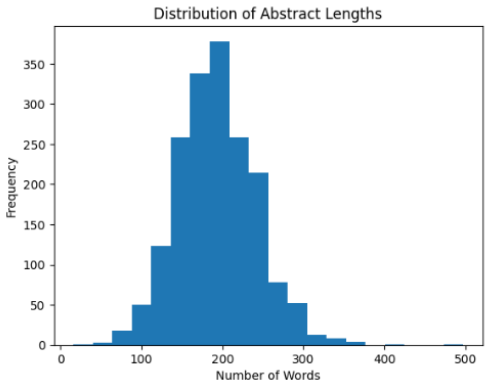
dengan cara mengalokasikan setiap titik data ke centroid yang paling dekat berdasarkan perhitungan jarak Euclidean, kemudian secara berulang memutakhirkan posisi setiap centroid hingga proses konvergensi tercapai dan tidak terjadi lagi perubahan signifikan pada posisi centroid. Sasaran utama yang ingin dicapai oleh algoritma ini adalah menekan seminimal mungkin total kuadrat jarak antara setiap titik data dengan centroid klaster yang menaunginya (*sum of squared errors*)[6]. Penetapan jumlah klaster optimal (k) dalam penelitian ini dilakukan melalui serangkaian pengujian terhadap berbagai variasi nilai k, di mana kualitas setiap konfigurasi klaster yang dihasilkan dievaluasi menggunakan metrik *Silhouette Score*. Nilai k yang menghasilkan rata-rata skor tertinggi selanjutnya ditetapkan sebagai jumlah klaster yang paling optimal, yakni k = 4. Pada tahap awal proses pengelompokan, posisi centroid diinisialisasi secara acak sebagai titik awal, kemudian diperbarui secara berulang melalui proses iteratif sampai tidak ditemukan lagi pergeseran yang berarti pada posisi masing-masing centroid. Adapun jarak antara dua titik data dalam ruang fitur dikalkulasi menggunakan formula *Euclidean distance* sebagai berikut:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (6)$$

Dalam formula tersebut, *x* dan *y* merepresentasikan dua buah vektor data yang berada dalam ruang berdimensi-n. Nilai jarak yang dihitung melalui formula ini dimanfaatkan sebagai ukuran kedekatan antar dokumen, di mana semakin kecil nilai jarak yang dihasilkan, semakin erat pula tingkat kemiripan antara kedua dokumen yang dibandingkan.

HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini terdiri atas 1.200 abstrak artikel ilmiah yang bersumber dari portal *Science and Technology Index* (SINTA). Proses pengumpulan data mencakup keenam tingkatan akreditasi jurnal yang tersedia dalam portal tersebut, mulai dari SINTA 1 sampai dengan SINTA 6, dengan jumlah abstrak yang dihimpun dari masing-masing tingkatan akreditasi sebanyak 300 artikel. Proses pengambilan data dilakukan menggunakan kata kunci “informatic” dan “computer” untuk memastikan kesesuaian data dengan ruang lingkup penelitian. Selain itu, hanya abstrak artikel yang lengkap, relevan, dan tidak duplikat yang digunakan sebagai dataset penelitian. Untuk memberikan gambaran karakteristik dataset secara menyeluruh, dilakukan analisis statistik terhadap data abstrak artikel. Hasil analisis menunjukkan bahwa rata-rata panjang abstrak berada pada kisaran 190–200 kata, dengan sebagian besar abstrak memiliki panjang antara 150 hingga 250 kata. Distribusi panjang abstrak menunjukkan pola yang relatif normal sehingga menunjukkan konsistensi data pada keseluruhan dataset. Selain itu, analisis kata yang paling sering muncul menunjukkan bahwa teks mentah masih didominasi oleh kata umum atau stopword. Setelah melalui tahap preprocessing, kata-kata yang lebih relevan dan spesifik terhadap bidang informatika menjadi lebih dominan. Selanjutnya, pembobotan TF-IDF digunakan untuk mengidentifikasi kata-kata penting yang merepresentasikan karakteristik utama pada setiap dokumen, Ditunjukkan pada Gambar 4.



Gambar 4. Dataset

Case folding merupakan tahap awal dalam rangkaian prapemrosesan teks yang berfungsi mengubah keseluruhan karakter dalam teks menjadi bentuk huruf kecil (*lowercase*). Penerapan langkah ini dimaksudkan untuk menciptakan keseragaman dalam representasi kata, sekaligus meminimalkan inkonsistensi yang timbul akibat perbedaan penggunaan huruf kapital antar dokumen, sehingga proses analisis teks selanjutnya dapat menghasilkan keluaran yang lebih presisi dan berkualitas tinggi. Hasil penerapan proses *case folding* dapat dilihat pada Tabel 1.

Tabel 1. Case folding

No	Dataset Abstract	Case Folding
1	online shopping has recently shown a very rapid development and plus there is a deadly virus that makes people afraid to leave the house, so an alternative way to not have to leave the house is to shop online. with the passage of time, there has been a change in shopping, where you usually have to go to the place directly, now you don't have to go directly to the place of shopping because of the existing technology. so with more and more people buying things online, the e-commerce site must look good and be easy to use. this research was conducted to design a website-based e-commerce design to make it more attractive. making this research using data collection	online, shopping, has, recently, shown, a, very, rapid, development, and, plus, there, is, a, deadly, virus, that, makes, people, afraid, to, leave, the, house, so, an, alternative, way, to, not, have, to, leave, the, house, is, to, shop, online, with, the, passage, of, time, there, has, been, a, change, in, shopping, where, you, usually, have, to, go, to, the, place, directly, now, you, don, t, have, to, go, directly, to, the, place, of, shopping, because, of, the, existing, technology, so, with, more, and, more, people, buying, things, online, the, e, commerce, site, must, look, good, and, be, easy, to, use, this, research, was, conducted, to, design, a, website, based, e, commerce, design, to, make, it, more, attractive, making, this, research, using, data, collection
...	....	....
1200	Food hygiene is one of the important factors in maintaining health, however food hygiene has not fully been implemented in the community. Therefore, Artificial Intelligence technology is presented in Society 5.0 to improve food safety and quality. This study aims to analyze opportunities for utilization of AI technology in the food sector. In conducting research ranging from problem identification, standard search, data collection, data analysis, to conclusions and suggestions. The results of the study obtained the magnitude of opportunities to utilize AI technology in the food sector. AI technology can improve the safety of food hygiene and facilitate monitoring of environmental hygiene, workers, or food itself.	food hygiene is one of the important factors in maintaining health, however food hygiene has not fully been implemented in the community. therefore, artificial intelligence technology is presented in society 5.0 to improve food safety and quality. this study aims to analyze opportunities for utilization of ai technology in the food sector. in conducting research ranging from problem identification, standard search, data collection, data analysis, to conclusions and suggestions. the results of the study obtained the magnitude of opportunities to utilize ai technology in the food sector. ai technology can improve the safety of food hygiene and facilitate monitoring of environmental hygiene, workers, or food itself.

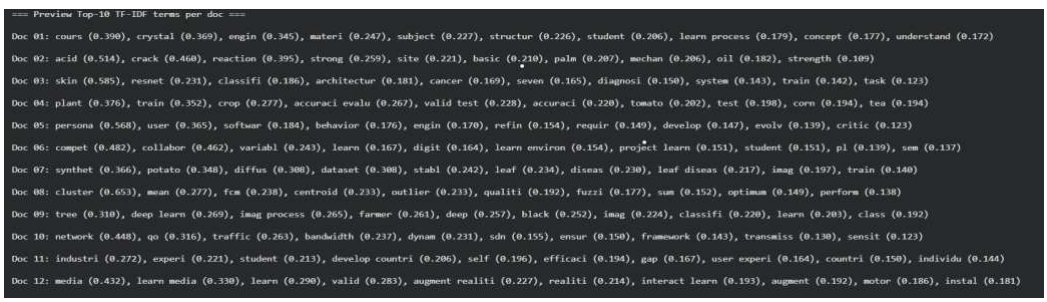
Tokenization adalah proses dalam tahap prapemrosesan yang bertujuan untuk membagi teks menjadi kata-kata individual (token) sehingga memudahkan proses analisis selanjutnya. Hasil dari tahapan ini disajikan pada Tabel 3. Setelah itu, dilakukan *stopword removal*, yaitu penghilangan kata-kata umum yang frekuensinya tinggi tetapi tidak memberikan kontribusi substantif dalam mengidentifikasi atau membedakan topik dokumen. Proses *stopword*

removal meningkatkan kualitas representasi teks dengan menghilangkan istilah-istilah yang tidak memiliki informasi penting. Sementara itu, *stemming* bertujuan untuk mengurangi variasi kata dengan mengubah kata berimbuhan menjadi bentuk dasarnya. Hasil dari kedua tahapan prapemrosesan tersebut disajikan pada Tabel 2.

Tabel 2. Prapemrosesan data

No	Tokenisasi	Stopword Removal	Stemming
1	online, shopping, has, recently, shown, a, very, rapid, development, and, plus, there, is, a, deadly, virus, that, makes, people, afraid, to, leave, the, house, so, an, alternative, way, to, not, have, to, leave, the, house, is, to, shop, online, with, the, passage, of, time, there, has, been, a, change, in, shopping, where, you, usually, have, to, go, to, the, place, directly, now, you, don, t, have, to, go, directly, to, the, place, of, shopping, because, of, the, existing, technology, so, with, more, and, more, people, buying, things, online, the, e, commerce, site, must, look, good, and, be, easy, to, use, this, research, was, conducted, to, design, a, website, based, e, commerce, design, to, make, it, more, attractive, making, this, research, using, data, collection	online, shopping, recently, shown, rapid, development, plus, deadly, virus, makes, people, afraid, leave, house, alternative, way, leave, house, shop, online, passage, time, change, shopping, usually, place, directly, don, directly, place, shopping, existing, people, buying, things, online, commerce, site, look, good, easy, conducted, design, website, commerce, design, make, attractive, making, collectio	onlin, shop, recent, shown, rapid, develop, plu, deadli, viru, make, peopl, afraid, leav, hous, altern, way, leav, hous, shop, onlin, passag, time, chang, shop, usual, place, directli, don, directli, place, shop, exist, peopl, buy, thing, onlin, commerc, site, look, good, easi, conduct, design, websit, commerc, design, make, attract, make, collec
...	... .	... .	... ...
1.200	online, shopping, has, recently, shown, a, very, rapid, development, and, plus, there, is, a, deadly, virus, that, makes, people, afraid, to, leave, the, house, so, an, alternative, way, to, not, have, to, leave, the, house, is, to, shop, online, with, the, passage, of, time, there, has, been, a, change, in, shopping, where, you, usually, have, to, go, to, the, place, directly, now, you, don, t, have, to, go, directly, to, the, place, of, shopping, because, of, the, existing, technology, so, with, more, and, more, people, buying, things, online, the, e, commerce, site, must, look, good, and, be, easy, to, use, this, research, was, conducted, to, design, a, website, based, e, commerce, design, to, make, it, more, attractive, making, this, research, using, data, collection	food, hygiene, important, factors, maintaining, health, food, hygiene, fully, implemented, community, artificial, intelligence, presented, society, improve, food, safety, quality, aims, analyze, opportunities, utilization, food, sector, conducting, ranging, problem, identification, standard, search, collection, conclusions, suggestions, obtained, magnitude, opportunities, utilize, food, sector, improve, safety, food, hygiene, facilitate, monitoring, environmental, hygiene, workers, food	food, hygien, import, factor, maintain, health, food, hygien, fulli, implement, commun, artifici, intellig, present, societi, improv, food, safeti, qualiti, aim, analyz, opportun, util, food, sector, conduct, rang, problem, identif, standard, search, collect, conclus, suggest, obtain, magnitud, opportun, util, food, sector, improv, safeti, food, hygien, facilit, monitor, environment, hygien, worker, food

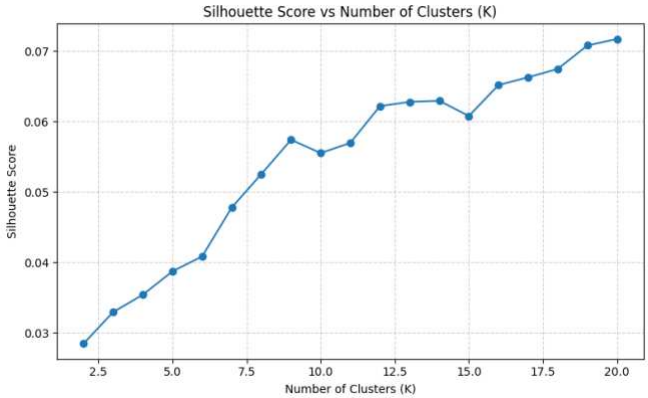
Hasil ekstraksi fitur menggunakan metode TF-IDF menunjukkan bahwa beberapa kata dominan seperti *learning*, *system*, *algorithm*, dan *accuracy* banyak muncul pada abstrak artikel, yang mencerminkan fokus penelitian pada pengembangan metode dan performa sistem. Pembobotan TF-IDF mampu mengidentifikasi kata-kata penting yang merepresentasikan karakteristik utama dokumen sehingga efektif digunakan dalam representasi fitur teks. Hasil pembobotan kata menggunakan metode TF-IDF ditunjukkan pada Gambar 5.



Gambar 5. Hasil TF-IDF

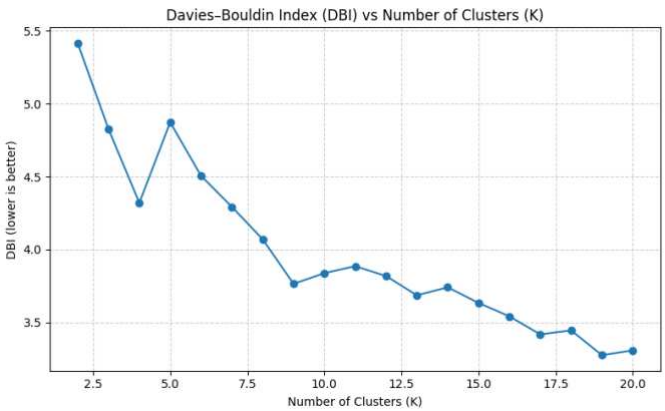
Fitur TF-IDF yang dihasilkan kemudian digunakan sebagai masukan pada proses clustering menggunakan algoritma K-Means. Sebelum proses clustering dilakukan, fitur TF-IDF direduksi dimensinya menggunakan TruncatedSVD dan dinormalisasi untuk meningkatkan efisiensi serta stabilitas proses clustering. Selanjutnya, algoritma K-Means diterapkan pada ruang fitur hasil reduksi untuk mengelompokkan dokumen berdasarkan kemiripan kata menggunakan Euclidean Distance.

Grafik *Silhouette Score vs Number of Clusters (K)* menunjukkan evaluasi kualitas klaster menggunakan metode Silhouette Score pada berbagai jumlah klaster. Berdasarkan grafik, nilai Silhouette Score cenderung meningkat seiring bertambahnya jumlah klaster, yang menandakan bahwa struktur klaster menjadi lebih baik dan pemisahan antar klaster semakin jelas. Nilai tertinggi diperoleh pada K = 20 dengan Silhouette Score sekitar 0,071, sehingga jumlah klaster tersebut dianggap memberikan hasil clustering terbaik pada dataset penelitian ini. Namun, secara umum nilai Silhouette Score yang relatif rendah menunjukkan bahwa tingkat separasi antar klaster masih belum terlalu kuat.pada Gambar 6.



Gambar 6. *Silhouette*

Grafik *Davies-Bouldin Index (DBI) vs Number of Clusters (K)* menunjukkan bahwa nilai DBI cenderung menurun seiring bertambahnya jumlah klaster. Penurunan nilai DBI mengindikasikan bahwa kualitas clustering semakin baik karena jarak antar klaster menjadi lebih jelas dan tingkat kemiripan dalam klaster semakin tinggi. Nilai DBI terendah diperoleh pada K = 19 dengan nilai sekitar 3,27, sehingga jumlah klaster tersebut dianggap memberikan performa clustering terbaik berdasarkan evaluasi DBI.



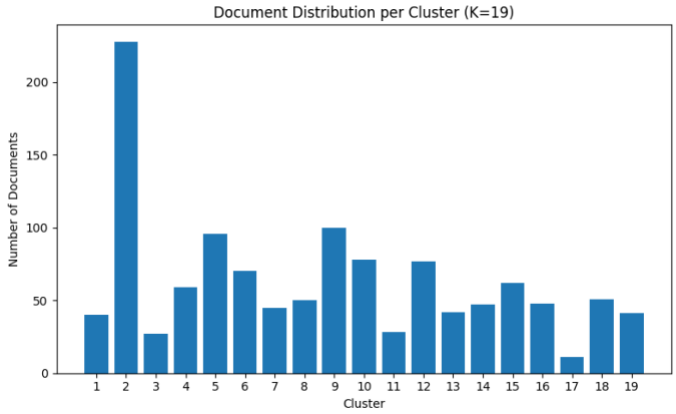
Gambar 7. *DBI*

Tabel evaluasi menunjukkan hasil pengukuran kualitas clustering menggunakan metode Silhouette Score dan Davies–Bouldin Index (DBI) pada berbagai jumlah cluster (K). Berdasarkan hasil tersebut, nilai Silhouette Score cenderung meningkat seiring bertambahnya jumlah cluster, sedangkan nilai DBI cenderung menurun. Hal ini menunjukkan bahwa kualitas clustering menjadi semakin baik. Nilai Silhouette Score tertinggi diperoleh pada K = 20 sebesar 0,071682, sedangkan nilai DBI terendah diperoleh pada K = 19 sebesar 3,276018. Dengan demikian, K = 19 dan K = 20 dapat dianggap sebagai jumlah cluster yang memberikan performa clustering terbaik pada penelitian ini, Tabel 4 dibawah ini.

Tabel 3 evalusi Silhouete dan DBI

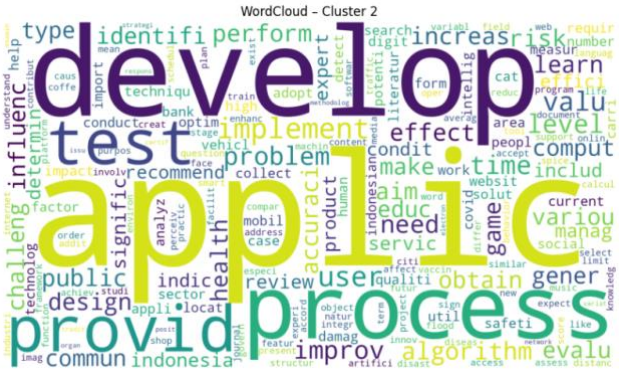
No	Kluster	Silhouete	Davies–Bouldin Index
0	K = 2	0.028448	5.416482
1	K = 3	0.032928	4.826318
2	K = 4	0.035390	4.320467
3	K = 5	0.038711	4.873131
4	K= 6	0.040818	4.505837
5	K= 7	0.047766	4.293196
6	K= 8	0.052516	4.070352
7	K= 9	0.057379	3.765164
8	K=10	0.055498	3.838298
11	……	……	…….
17	K=19	0.070766	3.276018
18	K=20	0.071682	3.308686

Berdasarkan hasil evaluasi clustering, algoritma K-Means diterapkan menggunakan jumlah cluster optimal, yaitu K=19. Proses clustering mencapai konvergensi setelah 18 iterasi, yang menunjukkan terbentuknya cluster yang stabil dari total 1.200 dokumen. Nilai *Silhouette Score* sebesar 0,0708 menunjukkan bahwa pemisahan antar cluster belum terlalu kuat, yang merupakan karakteristik umum pada data teks berdimensi tinggi akibat representasi TF-IDF yang *sparse* dan adanya kemiripan kosakata antar topik penelitian. Meskipun demikian, hasil clustering tetap mampu merepresentasikan struktur tematik dataset, didukung oleh nilai *Davies-Bouldin Index* (DBI) sebesar 3,2760 yang menunjukkan tingkat kekompakan dan pemisahan cluster yang cukup baik. Gambar 8 menunjukkan distribusi dokumen pada masing-masing cluster yang memiliki jumlah bervariasi. Cluster 2 menjadi cluster dengan jumlah dokumen terbanyak, yaitu 228 dokumen, diikuti *Cluster* 9 sebanyak 100 dokumen dan *Cluster* 5 sebanyak 96 dokumen, yang mengindikasikan bahwa topik pada cluster tersebut lebih dominan dalam dataset. Sebaliknya, *Cluster* 17 memiliki jumlah dokumen paling sedikit, yaitu 11 dokumen, yang menunjukkan topik yang lebih spesifik atau jarang dibahas. Variasi distribusi dokumen ini menunjukkan bahwa proses *clustering* berhasil mengelompokkan dokumen berdasarkan kesamaan topik dan karakteristik istilah yang muncul pada setiap dokumen.



Gambar 8. *clustering*

Hasil *clustering* menghasilkan 19 klaster yang merepresentasikan berbagai topik penelitian berdasarkan kemunculan kata-kata dominan pada masing-masing word cloud. Secara umum, klaster yang terbentuk mencerminkan tema dalam bidang teknologi informasi dan komputasi, seperti machine learning, analisis sentimen, jaringan komputer, *Internet of Things* (IoT), teknologi pendidikan, serta pengolahan data. Berdasarkan distribusi dokumen pada setiap klaster, dipilih beberapa klaster representatif untuk divisualisasikan, yaitu klaster dengan jumlah dokumen dominan, menengah, dan menengah-bawah. Klaster 2 dipilih karena memiliki jumlah dokumen paling besar, diikuti oleh klaster 9 sebagai representasi klaster dengan jumlah dokumen menengah, serta klaster 19 yang merepresentasikan klaster dengan jumlah dokumen menengah-bawah. Pemilihan ini bertujuan untuk memberikan gambaran variasi karakteristik topik yang dihasilkan dari proses *clustering*. Berdasarkan hasil visualisasi, kemunculan kata seperti learn, model, data, dan accuracy mengindikasikan fokus pada pengembangan model pembelajaran mesin serta evaluasi kinerja sistem, sementara kata *network, packet*, dan internet menunjukkan pembahasan terkait jaringan komputer dan komunikasi data. Selain itu, kata *student*, *teacher*, dan *learn* mencerminkan topik teknologi pendidikan, sedangkan kata sentiment, tweet, dan review mengarah pada analisis sentimen pada media sosial, serta kata sensor, *temperature*, dan monitoring mengindikasikan penerapan IoT dalam sistem pemantauan. Dengan demikian, hasil clustering menunjukkan bahwa dokumen-dokumen yang dianalisis cenderung terkelompok berdasarkan kesamaan topik dan karakteristik istilah yang muncul, meskipun interpretasi yang dihasilkan bersifat indikatif.



Gambar 8. *WordCloud*

Gambar 2. *WordCloud*

KESIMPULAN

Penelitian ini telah berhasil mengimplementasikan algoritma *K-Means Clustering* sebagai pendekatan pengelompokan abstrak artikel jurnal di bidang informatika berdasarkan kedekatan topik, dengan memanfaatkan representasi fitur berbasis TF-IDF sebagai fondasi pembobotan teks. Serangkaian tahapan prapemrosesan yang diterapkan, mencakup *case folding*, tokenisasi, *stopword removal*, dan *stemming*, terbukti mampu menghasilkan representasi dokumen yang lebih rapi dan terorganisir, sehingga memberikan dukungan yang signifikan terhadap efektivitas proses pengelompokan dokumen secara keseluruhan. Berdasarkan hasil pengukuran menggunakan dua metrik evaluasi, yakni *Silhouette Score* dan *Davies–Bouldin Index* (DBI), konfigurasi klaster yang paling optimal ditemukan pada K = 19, dengan perolehan nilai *Silhouette Score* sebesar 0,0708 dan nilai DBI sebesar 3,2760. Nilai *Silhouette Score* yang terbilang rendah mengindikasikan bahwa tingkat keterpisahan antar klaster yang terbentuk masih belum cukup kuat, namun demikian kondisi semacam ini merupakan fenomena yang lazim dijumpai pada data teks berdimensi tinggi yang memiliki karakteristik topik saling tumpang tindih satu sama lain. Di sisi lain, nilai DBI yang diperoleh mencerminkan bahwa model yang dibangun mampu mencapai tingkat kekompakan internal klaster dan ketegasan batas antar klaster yang cukup memadai. Temuan ini selaras dengan sejumlah kajian terdahulu yang menegaskan bahwa integrasi TF-IDF dan K-Means merupakan kombinasi yang andal dalam menangani permasalahan pengelompokan data teks, kendati metrik evaluasi internal seperti *Silhouette Score* dan DBI pada data bertipe berdimensi tinggi memang cenderung tidak menghasilkan nilai yang tinggi [24].

Apabila disandingkan dengan penelitian-penelitian terdahulu yang memusatkan perhatian pada analisis topik dokumen dan pengukuran kualitas pengelompokan menggunakan pendekatan serupa, penelitian ini menawarkan kontribusi yang lebih spesifik, yakni pengelompokan abstrak artikel jurnal di bidang informatika yang bersumber dari portal SINTA pada rentang periode 2023–2024, dengan cakupan dataset yang lebih terdefinisi dan tersusun secara sistematis. Ditinjau secara menyeluruh, sinergi antara metode representasi TF-IDF dan algoritma K-Means telah terbukti memberikan dukungan yang memadai dalam mengorganisasikan abstrak artikel jurnal informatika ke dalam kelompok-kelompok berdasarkan kedekatan topiknya. Untuk penelitian yang akan datang, disarankan agar dikembangkan pendekatan representasi fitur yang lebih kaya secara semantik, seperti pemanfaatan LDA, Word2Vec maupun BERT, serta mengeksplorasi penerapan algoritma pengelompokan alternatif yang berpotensi menghasilkan pemisahan antar klaster dengan kualitas yang lebih unggul.

DAFTAR PUSTAKA

[1] R. F. S. Cahyanto, A. R. Chrismanto, and D. Sebastián, "Pengelompokan Komentar Dataset Sentipol Dengan Modified K-Means Clustering," *Jurnal Teknik Informatika Dan Sistem Informasi*, 2020, doi: 10.28932/jutisi.v6i3.3006.

[2] E. Delibaş, "Efficient TF-IDF method for alignment-free DNA sequence similarity analysis," *J. Mol. Graph. Model.*, vol. 137, Jun. 2025, doi: 10.1016/j.jmgm.2025.109011.

[3] P. Guleria, J. Frnda, and P. N. Srinivasu, "NLP based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis," *Array*, vol. 27, Sep. 2025, doi: 10.1016/j.array.2025.100467.

[4] J. P. Pamput, A. R. Muthmainnah, A. A. N. Risal, and D. F. Surianto, "K-Means++ and TF-IDF for Grouping Library Books by Topic," *Paradigma - Jurnal Komputer dan Informatika*, vol. 27, no. 2, pp. 74–82, Sep. 2025, doi: 10.31294/p.v27i2.8272.

[5] I. M. Nur and Abdurakhman, "K-Means ++ Algorithm for Health Services Clustering Based on Districts in West Java Province," *EKSAKTA: Journal of Sciences and Data Analysis*, pp. 96–102, Apr. 2024, doi: 10.20885/eksakta.vol5.iss1.art11.

[6] Muhammad Raqib Syahkur, D. Hartama, and S. Solikhun, "Evaluasi Jumlah Cluster pada Algoritma K-Means++ Menggunakan Silhouette dan Elbow dengan Validasi Nilai DBI dalam Mengelompokkan Gizi Balita," *JST (Jurnal Sains dan Teknologi)*, vol. 13, no. 3, pp. 487–496, Oct. 2024, doi: 10.23887/jstundiksha.v13i3.86419.

[7] D. F. Surianto and D. F. Surianto, "Enhancing K-Means Clustering for Journal Articles using TF-IDF and LDA Feature Extraction," *Brilliance: Research of Artificial Intelligence*, vol. 4, no. 2, pp. 964–972, Mar. 2025, doi: 10.47709/brilliance.v4i2.5547.

[8] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, Dec. 2019, doi: 10.1186/s13673-019-0192-7.

[9] S. M. Al-Ghuribi, S. A. M. Noah, M. A. Mohammed, S. N. Qasem, and B. A. H. Murshed, "To Cluster or Not to Cluster: The Impact of Clustering on the Performance of Aspect-Based Collaborative Filtering," *Ieee Access*, 2023, doi: 10.1109/access.2023.3270260.

[10] D. Rifaldi, A. Fadlil, and Herman, "Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet 'Mental Health,'" *Decode Jurnal Pendidikan Teknologi Informasi*, 2023, doi: 10.51454/decode.v3i2.131.

[11] A. Suparto, M. J. Clement, A. M. A. Rahman, and H. Irsyad, "Implementasi Preprocessing Dan Synonym Expansion Untuk Sistem Temu Kembali Berita Bahasa Indonesia," 2025, doi: 10.36982/jseci.v3i01.5405.

[12] I. A. Putri, F. Fadiana, and D. Y. Kristianto, "Pemetaan Tempat Ibadah Di Jakarta: Sebuah Studi Menggunakan Kmeans Clustering," *Jurnal Rekayasa Sistem Informasi Dan Teknologi*, 2024, doi: 10.59407/jrsit.v1i4.669.

[13] J. N. Iin, "Perbandingan Kmeans Dan Hierarchical Clustering Untuk Pemetaan Kawasan Rawan Stunting Di Kabupaten Kolaka," *Rabit Jurnal Teknologi Dan Sistem Informasi Univrab*, 2025, doi: 10.36341/rabit.v10i2.6358.

[14] B. Priandini, M. Asfi, and L. Magdalena, "Implementasi K-Means Rfm Dan Holt-Winters Exponential Smoothing Additive Dalam Sistem Business Intelligence Untuk Strategi Pengelolaan Pelanggan Pada Perusahaan Transportasi. Pembuatan Dashboard BI Segmentasi Pelanggan Dan Peramalan Jumlah Pelanggan Menggunakan Tools Tableau Menggunakan Metode Kmeans RFM Dan Holtwinters Exponential Smoothing," *Methodika Jurnal Teknik Informatika Dan Sistem Informasi*, 2025, doi: 10.46880/mtk.v11i2.4511.

[15] K. Kelvin, F. M. Sinaga, Wulan Sri Lestari, Sunaryo Winardi, K. Hawani Rambe, and R. Purba, "Enhancing Sentiment Analysis Accuracy With Bert And Silhouette Method Optimization," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 11, no. 1, Aug. 2025, doi: 10.33480/jitk.v11i1.6392.

[16] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, Jun. 2021, doi: 10.3390/e23060759.

[17] 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 Jurnal Ilmiah Teknologi Dan Komputer*, vol. 3, no. 1, p. 821, 2022, doi: 10.24843/jtrti.2022.v03.i01.p18.

[18] 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 Jurnal Ilmiah Teknologi Dan Komputer*, vol. 3, no. 1, p. 821, 2022, doi: 10.24843/jtrti.2022.v03.i01.p18.

[19] R. H. Astuti, M. Muljono, and S. Sutriawan, "Indonesian News Text Summarization Using MBART Algorithm," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 155–164, Feb. 2024, doi: 10.15294/sji.v11i1.49224.

[20] K. Wang, Y. Ding, and S. C. Han, "Graph Neural Networks for Text Classification: A Survey," *Artif. Intell. Rev.*, 2024, doi: 10.1007/s10462-024-10808-0.

[21] R. F. Salsabila, D. D. Prasetya, T. Widyaningtyas, and T. Hirashima, "Comparison of Text Representation for Clustering Student Concept Maps," *Matrik Jurnal Manajemen Teknik Informatika Dan Rekayasa Komputer*, vol. 24, no. 2, pp. 259–272, 2025, doi: 10.30812/matrik.v24i2.4598.

[22] W. L. Azzahra and E. Mailoa, "Analisis Sentimen Terhadap RSUD Salatiga Menggunakan SVM Dan TF-IDF," *Jurnal Indonesia Manajemen Informatika Dan Komunikasi*, 2025, doi: 10.35870/jimik.v6i1.1208.

[23] B. N. Yulisasih, H. Herman, S. Sunardi, and H. Yuliansyah, "Evaluation of K-Means Clustering Using Silhouette Score Method on Customer Segmentation," *ILKOM Jurnal Ilmiah*, vol. 16, no. 3, pp. 330–342, Dec. 2024, doi: 10.33096/ilkom.v16i3.2325.330-342.

[24] N. B. Nisa, R. P. Nanda, Z. H. Kudadiri, B. A. Alfahri, and Mhd. Furqan, "Klasifikasi Berita detik.com Terkait Teknologi Informasi Menggunakan TF-IDF Dan Naive Bayes," *Jurnal Nasional Komputasi Dan Teknologi Informasi (Jnkti)*, 2025, doi: 10.32672/jnkti.v8i3.9171.