JSAI : Journal Scientific and Applied Informatics
Vol. 08, No. 3, November 2025, hal. 663~668
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Pemodelan Topik Berdasarkan Dokumen Penelitian Bidang Ilmu Komputer Menggunakan *Text Mining*

[a]Bakhtiar K, [b]Azhar Andika Putra, [c]Muhammad Al Hapiz, [d]Firga Abel Astiawan
Fakultas Ilmu Komputer, Universitas Sjakhyakirti, Palembang, Indonesia
[a]bakhtiarkaseem@gmail.com, [b]azharandikaputra@unisti.ac.id, [c]21610014@student.unisti.ac.id, [d]21620007@student.unisti.ac.id

Article Info

Article history:

Received, 2025-11-29
Revised, 2025-11-08
Accepted, 2025-11-11

Kata Kunci:

klasterisasi dokumen;
abstrak penelitian;
IndoBERT;
K-Means;
TF-IDF

ABSTRAK

Penelitian ini bertujuan untuk mengembangkan model klasterisasi dokumen menggunakan kombinasi model IndoBERT dan algoritma K-Means untuk mengelompokkan abstrak penelitian dalam bidang ilmu komputer dan teknologi. Data yang digunakan terdiri dari 1000 abstrak penelitian yang dibagi menjadi dua yaitu 80% untuk data pelatihan (800 abstrak) dan 20% untuk data pengujian (200 abstrak). Model IndoBERT digunakan untuk merepresentasikan abstrak menjadi vektor embedding yang kemudian diproses dengan algoritma K-Means untuk membentuk 10 klaster topik yaitu kecerdasan buatan, sistem dan jaringan komputer, pemrograman, keamanan siber dan lainnya. Tahap eksperimen pelatihan menggunakan data pelatihan untuk menghasilkan klaster dan *centroid* untuk memetakan dokumen baru ke dalam klaster yang sesuai. Evaluasi dilakukan dengan menggunakan beberapa metrik, yaitu akurasi, homogenitas klaster, *davies-bouldin index*, dan *silhouette score*. Hasil pengujian menunjukkan bahwa model yang dikembangkan memiliki akurasi sebesar 85%, yang menunjukkan kinerja yang baik dalam mengklasterkan data uji. Nilai homogenitas klaster sebesar 0.90 mengindikasikan bahwa dokumen yang seharusnya berada dalam klaster yang sama dikelompokkan dengan baik. Nilai *davies-bouldin index* sebesar 0.34 sedangkan nilai Silhouette Score sebesar 0.76.

Keywords:

Document clustering;
Research abstract;
IndoBERT;
K-Means;
TF-IDF

ABSTRACT

This study aimed to develop a document clustering model using a combination of the IndoBERT model and the K-Means algorithm to group research abstracts in the field of computer science and technology. The data used consisted of 1000 research abstracts, divided into two parts: 80% for training data (800 abstracts) and 20% for testing data (200 abstracts). The IndoBERT model was used to represent the abstracts as embedding vectors, which were then processed with the K-Means algorithm to form 10 topic clusters, including artificial intelligence, computer systems and networks, programming, cybersecurity, and others. The training experiment used the training data to generate clusters and centroids for mapping new documents into the appropriate clusters. Evaluation was carried out using several metrics, including accuracy, cluster homogeneity, Davies-Bouldin Index, and Silhouette Score. The testing results showed that the developed model achieved an accuracy of 85%, indicating good performance in clustering the test data. The cluster homogeneity value of 0.90 indicated that documents that should belong to the same cluster were grouped together effectively. The Davies-Bouldin Index value was 0.34, while the Silhouette Score was 0.76.

Penulis Korespondensi:

Bakhtiar K,
Fakultas Ilmu Komputer
Universitas Sjakhyakirti, Palembang, Indonesia
Email: bakhtiarkaseem@gmail.com

1. PENDAHULUAN

Seiring dengan meningkatnya jumlah publikasi ilmiah dalam berbagai disiplin ilmu, kebutuhan untuk model pengolahan informasi dalam dokumen penelitian secara efisien menjadi semakin perlu untuk dikembangkan [1], [2], [3]. Volume data yang terus bertambah menuntut adanya pendekatan yang mampu menyederhanakan kompleksitas informasi dan menyajikan wawasan yang bermakna dari kumpulan teks dalam skala besar [4], [5], [6]. Salah satu pendekatan yang berkembang pesat untuk menangani tantangan ini adalah *topic modeling*, yaitu teknik yang digunakan untuk mengidentifikasi dan mengelompokkan topik-topik tersembunyi dalam kumpulan dokumen berbasis teks. Pendekatan ini sangat penting dalam mendukung pengembangan sistem rekomendasi, pencarian dokumen yang relevan, serta pemetaan perkembangan ilmu pengetahuan dalam suatu bidang kajian [7], [8], [9].

Namun, untuk memperoleh hasil pemodelan topik yang optimal diperlukan mekanisme pemilihan parameter dan klasterisasi yang sesuai. Penelitian oleh Aziz et al. (2022) menggunakan kombinasi *term frequency inverse document frequency* (TF-IDF) dan *latent dirichlet allocation* (LDA) untuk mengelompokkan topik dokumen berbahasa Inggris yang berkaitan dengan bidang ilmu ekonomi [10]. Penelitian oleh Chen et al. (2024) menggabungkan *bidirectional encoder representations from transformers* (BERT) dan LDA untuk menganalisis data teks persepsi wisatawan dalam domain teks pariwisata [11]. Penelitian oleh Onan & Alhumyani (2024) mengembangkan FuzzyTP-BERT untuk meningkatkan pemodelan topik dan ekstraksi ringkasan dokumen [12]. Penelitian oleh Amiri & Karimianghadim (2024) mengadopsi pendekatan TF-IDF dan BERT serta LDA untuk klasterisasi topik dokumen ilmiah secara umum [13].

Penelitian oleh Mehta et al. (2021) mengimplementasikan kombinasi BERT dan K-Means untuk klasterisasi teks secara umum [14]. Penelitian oleh Weng et al. (2022) menerapkan BERT dan HDBSCAN untuk pengelompolan topik dalam publikasi ilmiah [15]. Penelitian oleh Marutho & Rustad (2024) mengintegrasikan TF-IDF, BERT untuk klasterisasi data teks secara umum [16]. Dari berbagai penelitian di atas, dapat disimpulkan bahwa penggunaan model BERT telah menjadi acuan sebagai representasi bahasa untuk menganalisis konteks semantik dari dataset teks. Ringkasan penelitian terkait menggambarkan pendekatan yang telah dikembangkan dalam pemodelan topik dokumen, khususnya dengan memanfaatkan kombinasi teknik representasi teks dan algoritma klasterisasi dapat dilihat pada **Tabel 1**.

Tabel 1 Penelitian Terkait

No	Sumber	*Vectorization*	*Language Modelling*	*Topic Clustering*	Topik	Bahasa
1	[10]	TF-IDF	x	LDA	Penelitian	Inggris
2	[11]	x	BERT	LDA	Pariwisata	Inggris
3	[12]	x	BERT	Fuzzy Topic Modelling	Umum	Inggris
4	[13]	TF-IDF	BERT	LDA	Penelitian	Inggris
5	[14]	x	BERT	K-Means	Umum	Inggris
6	[15]	x	BERT	HDBSCAN	Penelitian	Inggris
7	Usulan	TF-IDF	IndoBERT	K-Means	Penelitian Ilmu Komputer	Indonesia

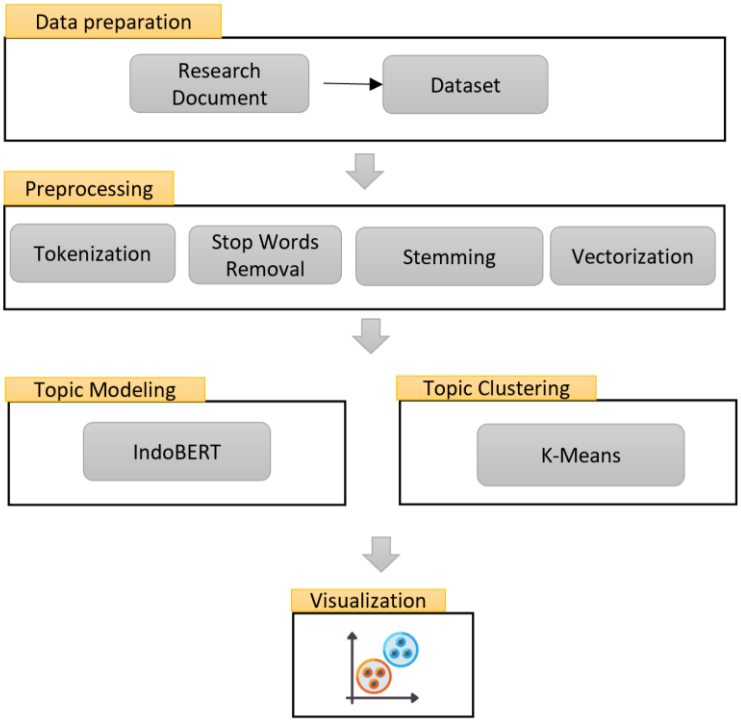
Pada penelitian yang diusulkan solusi masalah yang penelitian berupa integrasi antara representasi teks IndoBERT untuk bahasa Indonesia dan algoritma K-Means untuk proses klasterisasi topik. Berbeda dengan penelitian sebelumnya yang mayoritas menggunakan bahasa Inggris dan memanfaatkan model BERT dengan algoritma klasterisasi, penelitian ini secara khusus menggunakan dokumen penelitian berbahasa Indonesia.

Penelitian ini bertujuan untuk mengembangkan model pemodelan topik dokumen penelitian berbahasa Indonesia dengan mengintegrasikan representasi teks IndoBERT dan klasterisasi menggunakan K-Means. Dokumen penelitian berbahasa Indonesia terlebih dahulu diproses dan direpresentasikan dalam bentuk vector TF-IDF dan IndoBERT. Proses klasterisasi topik dilakukan menggunakan K-Means untuk mengelompokkan dokumen berdasarkan kemiripan kata berdasarkan klister topik penelitian ilmu komputer.

2. METODE PENELITIAN

Penelitian ini merupakan kelanjutan dari tahapan yang telah dilakukan pada tahun 2024, yaitu pengembangan aplikasi awal untuk manajemen dokumen penelitian. Aplikasi tersebut telah menyediakan fitur unggah, penyimpanan, serta kategorisasi manual berdasarkan metadata, sehingga menjadi fondasi bagi proses analisis lanjutan berbasis konten teks [17]. Pada tahun 2025, fokus penelitian diarahkan pada penerapan pemodelan topik secara otomatis dengan memanfaatkan representasi semantik IndoBERT dan proses klasterisasi topik menggunakan K-means. Metodologi penelitian ini dirancang untuk mengintegrasikan kemampuan pemahaman konteks bahasa Indonesia yang dihasilkan oleh IndoBERT dengan keunggulan algoritma metaheuristik dalam mengelompokkan dokumen berdasarkan kesamaan topik, sebagaimana digambarkan pada **Gambar 2**.

JSAI : Journal Scientific and Applied Informatics
Vol. 08, No. 3, November 2025, hal. 663~668
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085



Gambar 1 Metodologi Penelitian

Sumber: Diolah oleh peneliti, 2025

Penelitian ini memanfaatkan perangkat keras HP workstation high-performance yang dilengkapi dengan Graphics Processing Unit (GPU) NVIDIA Quadro RTX 4000. Dataset yang digunakan merupakan kumpulan dokumen ilmiah berbahasa Indonesia yang telah dikurasi dari berbagai sumber akademik dan publikasi nasional. Dokumen tersebut diperoleh dari repositori yang terdiri dari Garba Rujukan Digital (Garuda) yang dikelola oleh Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi (Kemendikbudristek), Science and Technology Index (SINTA) sebagai indeks jurnal nasional terakreditasi, serta Portal Jurnal Ilmiah yang dikelola oleh Badan Riset dan Inovasi Nasional (BRIN).

Metodologi penelitian ini terdiri dari beberapa tahapan utama untuk menghasilkan pemodelan topik dokumen penelitian berbahasa Indonesia. Pada tahap *data preparation* dimulai dengan pengumpulan dokumen penelitian yang relevan sebagai bahan analisis. Dokumen-dokumen ini dikompilasi dan diubah menjadi format dataset terstruktur sesuai atribut penelitian. Pada tahap pra-pemrosesan teks terdiri dari beberapa tahap utama, yaitu dengan *tokenization*, yaitu mengubah dokumen menjadi daftar token, seperti kata atau karakter. Selanjutnya, dilakukan *stop words removal* untuk menghilangkan kata-kata umum yang tidak memiliki makna. Tahap berikutnya adalah *stemming* dengan melakukan koversi kata-kata dikonversi ke bentuk dasar. Proses *vectorization* dilakukan dengan mengubah token menjadi bentuk numerik agar dapat digunakan dalam model pembelajaran mesin dengan menerapkan *term frequency-inverse document frequency* (TF-IDF) yang merepresentasikan bobot kata dalam dokumen berdasarkan frekuensi kemunculan dan sebarannya dalam keseluruhan korpus.

Representasi teks hasil pra-pemrosesan kemudian diproses menggunakan IndoBERT, model berbasis transformer yang dirancang untuk memahami konteks semantik dalam bahasa Indonesia. IndoBERT menghasilkan embedding kontekstual yang merepresentasikan makna dari dokumen penelitian. Vektor representasi dari IndoBERT kemudian dikelompokkan menggunakan K-Means untuk melakukan klasterisasi dan menentukan jumlah klaster pemodelan topik [18], [19], [20].

3. HASIL DAN ANALISIS

Dalam penelitian ini menggunakan pendekatan hybrid antara IndoBERT dan K-Means dengan tujuan klasterisasi teks. Data teks terlebih dahulu dinormalisasi ke dalam rentang nilai antara 0 hingga 1, dan populasi awal solusi dibentuk secara acak dalam batas tersebut. Setiap data tunggal dalam populasi merepresentasikan satu kemungkinan solusi klasterisasi dalam bentuk matriks [21] yang dipresentasikan seperti pada Persamaan (1).

JSAI : Journal Scientific and Applied Informatics
Vol. 08, No. 3, November 2025, hal. 663~668
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

$$x_i = (C_1, C_2, \dots, C_i, \dots, C_k) \quad (1)$$

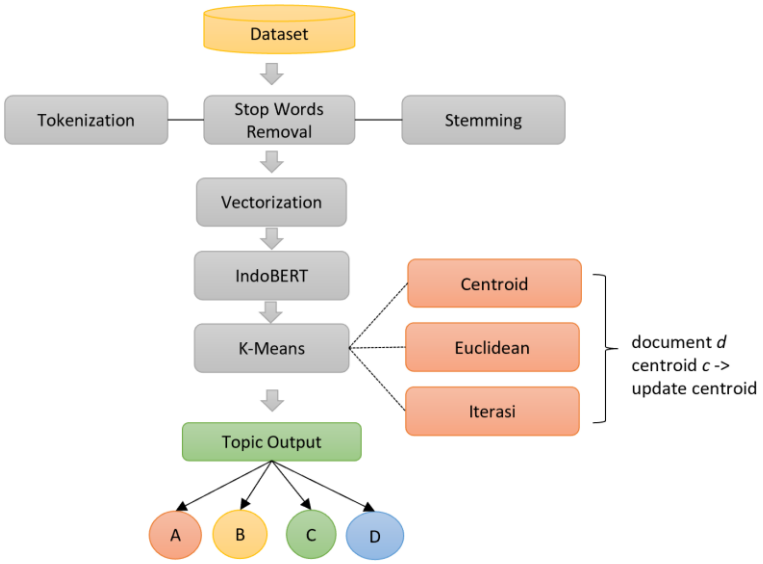
Nilai dari C_i adalah vector centroid dari klister ke-i dan $\mathcal{K}$ adalah jumlah total klister. Jarak antar dokumen d_j dan centroid C_k dihitung menggunakan rumus Euclidean [22] yang dipresentasikan seperti pada Persamaan (2).

$$dist(d_j, C_k) = \sqrt{\sum_{i=1}^{t} (d_{ij} - C_{ik})^2} \quad (2)$$

Dengan nilai t sebagai jumlah fitur dalam dokumen. Fungsi *fitness* dari setiap solusi dihitung berdasarkan rata-rata jarak dokumen terhadap centroid kluster masing-masing [22] yang dipresentasikan seperti pada Persamaan (3).

$$f(x) = \frac{1}{k} \sum_{i=1}^{k} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} dist\,(d_{ij}, C_i) \right) \quad (3)$$

Hasil klasterisasi divisualisasikan dalam bentuk grafik dua dimensi untuk menunjukkan distribusi dan keterkaitan antar topik yang terbentuk. Visualisasi ini mempermudah interpretasi hasil dan memberikan gambaran tentang struktur topik dalam kumpulan dokumen penelitian. Arsitektur dari pemodelan topik dari dokumen penelitian dapat dilihat pada **Gambar 2**.



Gambar 2 Arsitektur *Research Document Topic Modelling*
Sumber: Diolah oleh peneliti, 2025

Pada eksperimen penelitian, data yang digunakan terdiri dari 1000 abstrak penelitian, yang dibagi menjadi dua bagian yaitu data pelatihan dan data pengujian. Data pelatihan terdiri 80% dari total data, yaitu sebanyak 800 abstrak, yang digunakan untuk melatih model IndoBERT dalam menghasilkan vektor embedding dan melatih algoritma K-Means untuk mengklasterkan dokumen. Untuk presentase 20% yang terdiri dari 200 abstrak digunakan sebagai data pengujian untuk mengevaluasi performa model yang telah dilatih. Data pengujian ini bertujuan untuk memastikan bahwa model dapat mengklasterkan data baru dengan baik dan menghasilkan klaster yang sesuai dengan topik yang relevan.

Hasil klasterisasi dokumen penelitian menggunakan model IndoBERT dan K-Means terdiri dari klaster utama berkaitan topik dalam bidang ilmu komputer dan teknologi. Klaster pertama yaitu kecerdasan buatan berisi tentang machine learning dan aplikasi AI. Klaster sistem dan jaringan komputer membahas topik terkait jaringan komputer, arsitektur sistem, dan komunikasi data antar perangkat. Klaster pemrograman dan rekayasa perangkat lunak berisi tentang pengembangan perangkat lunak dan metodologi pemrograman. Klaster teori dan matematika komputer membahas penelitian teori komputasi dan algoritma komputer. Klaster sistem cerdas dan robotika berkaitan dengan riset robotika dan sistem cerdas sedangkan klaster pengolahan data dan *big data* berkaitan dengan pengelolaan dan analisis data dalam jumlah besar. Klaster keamanan siber merupakan abstrak penelitian tentang proteksi data dan keamanan sistem komputer. Klaster komputasi kuantum membahas tentang

teknologi komputasi kuantum dan klaster interaksi manusia-komputer merupakan abstrak peneliitan tentang interaksi antara manusia dengan komputer. Klaster etika dan dampak sosial teknologi membahas isu-isu etika dan dampak sosial dari penggunaan teknologi.

Pada tahap pelatihan, hasil klasterisasi menunjukkan bagaimana model IndoBERT dan K-Means meenggunakan 800 abstrak penelitian yang digunakan sebagai data pelatihan. Distribusi klaster yang dihasilkan berdasarkan klasterisasi K-Means selama proses pelatihan dapat dilihat pada **Tabel 2**.

Tabel 2 Hasil Data Pelatihan

Cluster	Topik	Jumlah Dokumen	Persentase
1	Kecerdasan Buatan	175	21.9%
2	Sistem dan Jaringan Komputer	140	17.5%
3	Pemrograman dan Rekayasa Perangkat Lunak	160	20%
4	Teori dan Matematika Komputer	60	7.5%
5	Sistem Cerdas dan Robotika	90	11.25%
6	Pengolahan Data dan Big Data	110	13.75%
7	Keamanan Siber	65	8.13%
8	Komputasi Kuantum	25	3.13%
9	Interaksi Manusia-Komputer	40	5%
10	Etika dan Dampak Sosial Teknologi	30	3.75%

Pada tahap pengujian, model yang sudah dilatih menggunakan IndoBERT dan K-Means diuji dengan menggunakan data pengujian terdiri dari 200 abstrak penelitian. Eksperimen pengujian pada data uji beruapa abstrak penelitian yang diproses menggunakan model IndoBERT untuk menghasilkan representasi vektor embedding, yang merupakan representasi numerik dari teks dalam ruang vektor. Nilai vektor tersebut dinormalisasi dalam rentang [0, 1] sesuai dengan model pelatihan. Selanjutnya, penerapan model K-Means dilakukan dengan menghitung jarak setiap vektor hasil representasi IndoBERT dari data pengujian ke centroid klaster yang telah terbentuk selama tahap pelatihan. Setiap dokumen pengujian kemudian dipetakan ke salah satu dari 10 klaster yang telah ada.

Untuk evaluasi hasil pengujian, digunakan beberapa metrik untuk mengukur kualitas klasterisasi, yaitu Homogenitas Klaster, Davies-Bouldin Index, dan Silhouette Score. Homogenitas Klaster digunakan untuk evaluasi kinerja berdasarkan dokumen yang seharusnya berada dalam klaster yang sama. Teknik Davies-Bouldin Index digunakan untuk evaluasi kinerja berdasarkan jarak antar klaster dibandingkan dengan jarak dalam klaster itu sendiri, dengan nilai yang lebih kecil menunjukkan klaster yang lebih terpisah dan lebih baik. Teknik Silhouette Score digunakan untuk evaluasi kinerja berdasarkan tingkat kepadatan dan keterpisahan klaster, memberikan representasi tentang dokumen yang ditempatkan dalam klaster yang sesuai. Hasil evaluasi menggunakan data pengujian dapat dilihat pada **Tabel 3**.

Tabel 3 Evaluasi Tahap Pengujian

No	Jenis Evaluasi	Nilai
1	Akurasi	85%
2	Homogenitas Klaster	0.90
3	Davies-Bouldin Index	0.34
4	Silhouette Score	0.76

4. KESIMPULAN

Kesimpulan dari penelitian ini membahas penggunaan model IndoBERT dan algoritma K-Means untuk klasterisasi dokumen penelitian dalam menghasilkan klaster topik-topik dalam bidang ilmu komputer dan teknologi. Berdasarkan data pelatihan, model berhasil membagi 800 abstrak penelitian ke dalam 10 klaster yang terdiri berbagai topik seperti kecerdasan buatan, sistem dan jaringan komputer, pemrograman, serta keamanan siber. Hasil evaluasi pada tahap pengujian dengan menggunakan 200 abstrak penelitian menunjukkan bahwa model yang telah dilatih memiliki kinerja akurasi mencapai 85% dan skor homogenitas klaster yang tinggi sebesar 0.90. Selain itu, nilai Davies-Bouldin Index sebesar 0.34 menunjukkan bahwa jarak antar klaster lebih besar dibandingkan dengan jarak dalam klaster, yang berarti klaste tersebut cukup terpisah dengan baik. Nilai Silhouette sebesar 0.76 menunjuukan bahwa klaster yang terbentuk memiliki tingkat kepadatan dan keterpisahan yang baik.

UCAPAN TERIMA KASIH

Tim peneliti berterima kasih kepada Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik Indonesia atas pendanaan hibah penelitian tahun 2025 dengan nomor 175/LL2/DT.05.00/PL/2025 dan 02/VII/PG.2/VI/2025 dan Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Sjakhyakirti atas dukungan yang telah diberikan untuk penyelesaian peneltiian.

REFERENSI

[1] J. D. Dinneen, M. Krtalić, N. Davoudi, H. Hellmich, C. Ochsner, and P. Bressel, "Information science and the inevitable: A literature review at the intersection of death and information management: An Annual Review of Information Science and Technology (ARIST) paper," *J. Assoc. Inf. Sci. Technol.*, vol. 75, no. 3, pp. 268–297, 2024.

[2] E. K. Donner, "Research data management systems and the organization of universities and research institutes: A systematic literature review," *J. Librariansh. Inf. Sci.*, vol. 55, no. 2, pp. 261–281, 2023.

[3] C. Gürkut, A. Elçi, and M. Nat, "An enriched decision-making satisfaction model for student information management systems," *Int. J. Inf. Manag. Data Insights*, vol. 3, no. 2, p. 100195, 2023.

[4] L. Yan and W. Zhiping, "Mapping the literature on academic publishing: A bibliometric analysis on WOS," *Sage Open*, vol. 13, no. 1, p. 21582440231158560, 2023.

[5] M.-J. Gallego-Losada, A. Montero-Navarro, E. García-Abajo, and R. Gallego-Losada, "Digital financial inclusion. Visualizing the academic literature," *Res. Int. Bus. Financ.*, vol. 64, p. 101862, 2023.

[6] A. Rejeb, K. Rejeb, A. Appolloni, Y. Kayikci, and M. Iranmanesh, "The landscape of public procurement research: a bibliometric analysis and topic modelling based on Scopus," *J. Public Procure.*, vol. 23, no. 2, pp. 145–178, 2023.

[7] V. Maphosa and M. Maphosa, "Artificial intelligence in higher education: a bibliometric analysis and topic modeling approach," *Appl. Artif. Intell.*, vol. 37, no. 1, p. 2261730, 2023.

[8] D. Yu and B. Xiang, "Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling," *Expert Syst. Appl.*, vol. 225, p. 120114, 2023.

[9] K. Kousha and M. Thelwall, "Artificial intelligence to support publishing and peer review: A summary and review," *Learn. Publ.*, vol. 37, no. 1, pp. 4–12, 2024.

[10] S. Aziz, M. Dowling, H. Hammami, and A. Piepenbrink, "Machine learning in finance: A topic modeling approach," *Eur. Financ. Manag.*, vol. 28, no. 3, pp. 744–770, 2022.

[11] Q. Chen, R. Liu, Q. Jiang, and S. Xu, "Exploring cross-cultural disparities in tourists' perceived images: a text mining and sentiment analysis study using LDA and BERT-BILSTM models," *Data Technol. Appl.*, vol. 58, no. 4, pp. 669–690, 2024.

[12] A. Onan and H. A. Alhumyani, "FuzzyTP-BERT: Enhancing extractive text summarization with fuzzy topic modeling and transformer networks," *J. King Saud Univ. Inf. Sci.*, vol. 36, no. 6, p. 102080, 2024.

[13] B. Amiri and R. Karimianghadim, "A novel text clustering model based on topic modelling and social network analysis," *Chaos, Solitons & Fractals*, vol. 181, p. 114633, 2024.

[14] V. Mehta, S. Bawa, and J. Singh, "WEClustering: word embeddings based text clustering technique for large datasets," *Complex Intell. Syst.*, vol. 7, no. 6, pp. 3211–3224, 2021.

[15] M.-H. Weng, S. Wu, and M. Dyer, "Identification and visualization of key topics in scientific publications with transformer-based language models and document clustering methods," *Appl. Sci.*, vol. 12, no. 21, p. 11220, 2022.

[16] D. Marutho and S. Rustad, "Optimizing aspect-based sentiment analysis using sentence embedding transformer, bayesian search clustering, and sparse attention mechanism," *J. Open Innov. Technol. Mark. Complex.*, vol. 10, no. 1, p. 100211, 2024.

[17] Bakhtiar. K, Lemi Iryani, and Renny.S, "Implementation of Laravel Framework and Waterfall SDLC Model for Research Document Information System Development," *JSAI (Journal Sci. Appl. Informatics)*, vol. 7, no. 2, pp. 405–410, Jun. 2024, doi: 10.36085/jsai.v7i2.7264.

[18] A. Petukhova, J. P. Matos-Carvalho, and N. Fachada, "Text clustering with large language model embeddings," *Int. J. Cogn. Comput. Eng.*, vol. 6, pp. 100–108, 2025.

[19] M. Hani'ah, V. N. Wijayaningrum, and A. R. Amalia, "Enhancing Extractive Summarization in Student Assignments Using BERT and K-Means Clustering," in *International Conference on Electronics, Biomedical Engineering, and Health Informatics*, Springer, 2023, pp. 453–464.

[20] J. Ravi and S. Kulkarni, "Text embedding techniques for efficient clustering of twitter data," *Evol. Intell.*, vol. 16, no. 5, pp. 1667–1677, 2023.

[21] W. Li and G.-G. Wang, "Improved elephant herding optimization using opposition-based learning and K-means clustering to solve numerical optimization problems," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 3, pp. 1753–1784, 2023.

[22] S. P. Haeri Boroujeni and E. Pashaei, "A hybrid chimp optimization algorithm and generalized normal distribution algorithm with opposition-based learning strategy for solving data clustering problems," *Iran J. Comput. Sci.*, vol. 7, no. 1, pp. 65–101, 2024.