Jurnal KomtekInfo

https://jkomtekinfo.org/ojs

2024 Vol. 11 No. 4 Hal: 332-341 e-ISSN: 2502-8758

Metode BERTopic dan LDA untuk Analisis Tren Penelitian Bidang Ilmu Komputer

Nursyahrina✉, Sarjon Defit, Rini Sovia

Fakultas Ilmu Komputer, Universitas Putra Indonesia YPTK, Padang, 25221, Indonesia

nursyahrina17@gmail.com

Abstract

Computer Science is a rapidly developing discipline, with the number of research publications increasing significantly in the last five years. Analysis of research trends in this field is still limited, so it is important to identify dominant research topics and understand the dynamics of their development. This study aims to analyze research topics and trends in the field of Computer Science using two topic modeling methods, namely Latent Dirichlet Allocation (LDA) and BERTopic. The data used consists of research article metadata obtained from the Emerald Insight website, with a total of 4,892 data in the 2019-2023 publication period. This study applies LDA and BERTopic to identify and group research topics based on title and abstract text. The embedding-based BERTopic method produces the highest coherence score of 0.49 in the model with a combination of TruncatedSVD-KMeans which identifies 13 topics, while LDA produces the highest coherence score of 0.42 in the model using the Bag-of-Words (BoW) feature extraction technique with 11 topics. The results of this study indicate that BERTopic is superior in producing more coherent and relevant topics compared to LDA, thanks to its ability to maintain semantic context between words in a document. This study is also able to provide important contributions to understanding the development of research trends in the field of Computer Science and can be a reference in planning future research.

Keywords: Computer Science, Research Trends, Topic Modeling, LDA, BERTopic.

Abstrak

Ilmu Komputer merupakan disiplin ilmu yang berkembang pesat, dengan jumlah publikasi penelitian yang meningkat secara signifikan dalam lima tahun terakhir. Analisis tren penelitian di bidang ini masih terbatas, sehingga penting untuk mengidentifikasi topik-topik penelitian dominan dan memahami dinamika perkembangannya. Penelitian ini bertujuan untuk menganalisis topik dan tren penelitian di bidang Ilmu Komputer dengan menggunakan dua metode *topic modeling*, yaitu *Latent Dirichlet Allocation* (LDA) dan BERTopic. Data yang digunakan terdiri dari metadata artikel penelitian yang diperoleh dari situs Emerald Insight, dengan total 4.892 data pada periode publikasi 2019-2023. Penelitian ini menerapkan LDA dan BERTopic untuk mengidentifikasi dan mengelompokkan topik-topik penelitian berdasarkan teks judul dan abstrak. Metode BERTopic yang berbasis *embedding* menghasilkan *coherence score* tertinggi sebesar 0,49 pada model dengan kombinasi TruncatedSVD-KMeans yang mengidentifikasi 13 topik, sementara LDA menghasilkan *coherence score* tertinggi sebesar 0,42 pada model yang menggunakan teknik ekstraksi fitur *Bag-of-Words* (BoW) dengan 11 topik. Hasil penelitian ini menunjukkan bahwa BERTopic lebih unggul dalam menghasilkan topik-topik yang lebih koheren dan relevan dibandingkan LDA, berkat kemampuannya dalam mempertahankan konteks semantik antar kata dalam dokumen. Penelitian ini juga mampu memberikan kontribusi penting dalam memahami perkembangan tren penelitian di bidang Ilmu Komputer dan dapat menjadi acuan dalam perencanaan penelitian di masa depan.

Kata kunci: Ilmu Komputer, Tren Penelitian, *Topic Modeling*, LDA, BERTopic.

1. Pendahuluan

Knowledge Discovery in Databases (KDD) muncul dari kebutuhan untuk menganalisis data dalam jumlah besar [1]. KDD adalah proses memperoleh pengetahuan dari data melalui penerapan metode-metode *data mining* [2]. Proses ini meliputi beberapa tahap, mulai dari seleksi dan prapemrosesan data, transformasi data, *data mining*, hingga interpretasi/evaluasi hasil, dengan *data mining* sebagai tahap utama. Dalam *data mining*, algoritma *machine learning* digunakan untuk menemukan pola dalam data [3],[4]. KDD tidak hanya terbatas pada data terstruktur seperti data numerik, tetapi juga menyediakan metode untuk menganalisis data tidak terstruktur seperti teks.

Text Mining, atau *Knowledge Discovery in Text Database*, berperan penting untuk menganalisis data teks dalam jumlah besar [1], menggunakan teknik-teknik *Natural Language Processing* (NLP) seperti *text preprocessing* [5] dan *feature extraction* [6], [7], [8]. Salah satu teknik penting dalam *text mining* adalah *topic modeling*, yaitu metode klasifikasi tanpa pengawasan (*unsupervised*) yang digunakan untuk mengidentifikasi struktur topik dalam koleksi teks dengan jumlah besar [9]. *Topic modeling* menggunakan algoritma seperti *Latent Dirichlet Allocation* (LDA) dan BERTopic untuk mengidentifikasi struktur topik dalam dokumen teks

Diterima: 28-08-2024 | Revisi: 17-10-2024 | Diterbitkan: 31-12-2024 | doi: 10.35134/komtekinfo.v11i4.580

besar, sering diterapkan dalam analisis tren penelitian [9], [10], [11].

Ilmu Komputer sebagai disiplin ilmu telah mengalami pertumbuhan pesat dalam hal publikasi penelitian, dengan lebih dari 2,9 juta publikasi tercatat dalam periode 2019-2023, menempatkannya di peringkat ketiga dalam kontribusi publikasi global setelah bidang Kedokteran dan Teknik [12]. Meskipun demikian, analisis tren penelitian dalam bidang ini masih kurang mendalam, yang menyulitkan pemahaman tentang arah perkembangan dan prospek topik penelitian di masa depan [13]. Untuk menangani keterbatasan ini, pendekatan modern seperti *topic modeling* menjadi penting dalam mengidentifikasi dan menganalisis tren penelitian. *Topic modeling*, dengan kemampuannya untuk mengelompokkan teks ke dalam topik-topik utama secara otomatis, dapat memberikan wawasan yang lebih mendalam tentang pola dan struktur dalam koleksi dokumen yang besar [9],[11].

Pada sebuah penelitian oleh Samsir dkk metode BERTopic digunakan untuk menganalisis 13.027 abstrak artikel Scopus tentang NLP. Penelitian ini menemukan 8 topik utama dengan nilai *coherence score* antara 0,42-0,58, menunjukkan bahwa BERTopic efektif dalam mengelompokkan dan mengidentifikasi literatur secara efisien. Temuan ini membuktikan bahwa BERTopic dapat menghasilkan topik yang koheren dan relevan dalam skala besar, memberikan wawasan yang mendalam tentang struktur penelitian di bidang NLP [14].

Penelitian lain oleh Akbarighatar dkk menerapkan BERTopic untuk mengevaluasi kapabilitas AI yang bertanggung jawab (*responsible* AI) dalam literatur terkait AI, menggunakan 1.451 abstrak artikel dari Scopus, Web of Science, dan *e-library* Association for Information Systems. Penelitian ini menemukan bahwa BERTopic efektif dalam menghasilkan enam tema utama yang relevan, menunjukkan kelebihan metode ini dalam analisis literatur yang kompleks. Hasil penelitian ini memberikan wawasan berharga dalam pengembangan kerangka kerja kapabilitas AI yang bertanggung jawab, menyoroti kekuatan BERTopic dalam menganalisis literatur secara menyeluruh [15].

Penelitian oleh Yu dkk menganalisis 33.957 artikel dari 25 jurnal terkemuka di domain fuzzy menggunakan LDA, yang mengidentifikasi 10 topik utama dan menunjukkan kemampuan LDA dalam mengungkap struktur pengetahuan serta perubahan tren penelitian di bidang *fuzzy* [16]. Sementara itu, Jung dan Kim menerapkan LDA untuk menganalisis 2.147 artikel dari jurnal SSCI dan SCIE tentang *sustainability* dan *marketing* dari tahun 2010 hingga 2020. Penelitian ini berhasil mengidentifikasi 14 topik laten, dengan topik lingkungan dan skenario keberlanjutan industri sebagai yang paling populer, yang menunjukkan efektivitas LDA dalam memetakan perubahan tren penelitian di bidang ini [17].

Perbandingan metode BERTopic dan LDA dilakukan oleh Kukushkin dkk pada penelitiannya dalam konteks *digital twins*, menggunakan 8.693 publikasi dari Scopus yang mencakup periode 1993-2022. Penelitian ini menunjukkan bahwa BERTopic menghasilkan lebih dari 100 topik, dengan 8 topik utama yang paling signifikan, sementara LDA menghasilkan 20 topik. Hasil ini menekankan keunggulan BERTopic dalam akurasi pengelompokan dan identifikasi topik yang relevan dibandingkan dengan LDA [18].

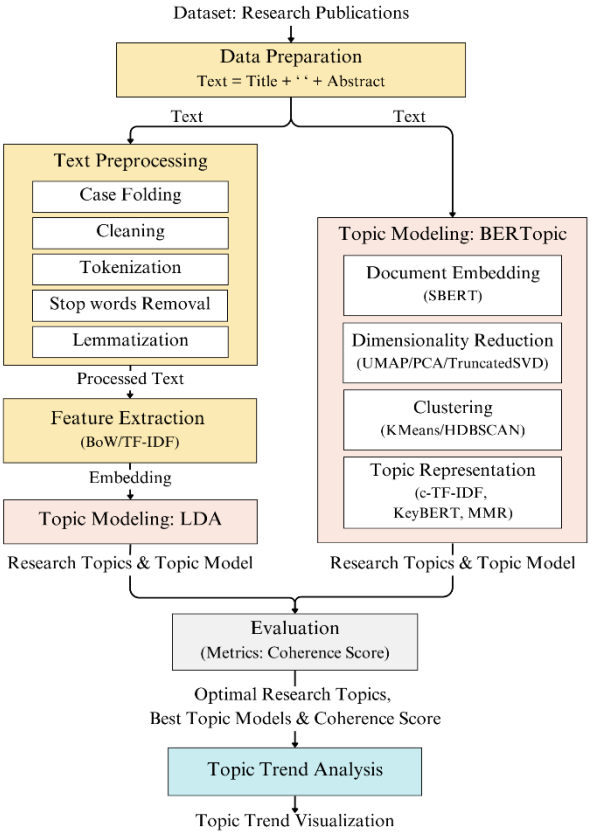
Penelitian lain oleh Ogunleye dkk juga membandingkan LDA, BERTopic, dan metode lainnya dalam analisis 10.000 *tweet* dari nasabah bank Nigeria. Mereka menemukan bahwa BERTopic dengan KernelPCA dan K-Means menghasilkan topik paling koheren dengan skor 0,8463, sementara LDA menunjukkan rentang koherensi antara 0,3-0,65. Penelitian ini memperkuat bukti bahwa BERTopic cenderung menghasilkan topik yang lebih berkualitas dan koheren dalam berbagai konteks analisis data tekstual [19].

Meskipun LDA dan BERTopic telah diterapkan dalam analisis topik di berbagai domain, penerapannya untuk analisis tren penelitian di bidang Ilmu Komputer masih terbatas. Penelitian sebelumnya cenderung menggunakan metode konvensional seperti analisis statistik bibliometrik yang kurang komprehensif dalam menangkap topik dan dinamika tren terkini [13], [20]. Penelitian ini bertujuan untuk mengisi kekurangan tersebut dengan menerapkan dan mengevaluasi BERTopic dan LDA dalam mengidentifikasi topik penelitian serta menganalisis dinamika tren di bidang Ilmu Komputer.

2. Metodologi Penelitian

Data penelitian merupakan metadata artikel jurnal terbitan 2019-2023, yang diunduh dari platform Emerald Insight. Pengumpulan data dilakukan melalui pencarian berbasis *query* dengan kata kunci terkait Ilmu Komputer dan 17 subbidangnya berdasarkan kurikulum CS2023 [21]. Jumlah awal data yang terkumpul adalah 6.000 artikel, yang kemudian digunakan untuk pengolahan dan analisis data sesuai langkah-langkah pada Gambar 1.

Tahapan analisis data dilakukan menggunakan program berbahasa Python. Program ini memanfaatkan *library* yang sudah ada dan umum digunakan, seperti pandas untuk pengolahan data dan regex serta nltk untuk *preprocessing*. *Library* lainnya termasuk gensim untuk *feature extraction*, *topic modeling* dengan LDA, dan evaluasi, bertopic dan sklearn untuk *topic modeling* dengan BERTopic, serta plotly untuk visualisasi data.



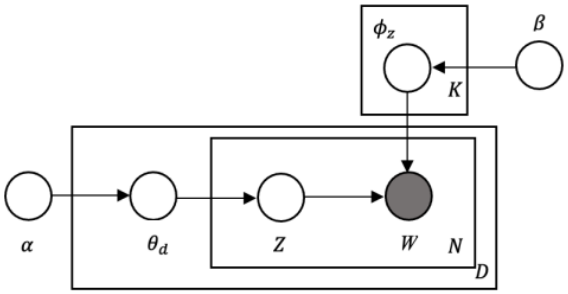
Gambar 1. Tahapan Analisis Data

Tahapan analisis dimulai dengan persiapan data, yang mencakup seleksi dokumen dalam Bahasa Inggris, pemilihan atribut penting seperti judul, abstrak, tahun publikasi, dan DOI, serta penghapusan data duplikat. Dari total 6.000 data awal, sebanyak 4.982 data berhasil diseleksi dan digunakan dalam penelitian ini. Selain itu, tahap persiapan data juga melibatkan penggabungan judul dan abstrak menjadi satu atribut teks baru yang diberi nama "*Text*" untuk mempermudah analisis lebih lanjut.

Selanjutnya, teks diproses melalui lima tahapan *text preprocessing—case folding*, *cleaning*, *tokenization*, *stop words removal*, dan *lemmatization*—yang diterapkan hanya pada LDA karena BERTopic dapat memproses data mentah secara langsung. Kemudian teks yang telah diproses diekstraksi menjadi vektor kata (*word embedding*) menggunakan teknik *feature extraction bag-of-words* (BOW) atau TF-IDF untuk LDA, sedangkan BERTopic menggunakan BERT *embedding* dalam proses pemodelan.

2.1. *Topic Modeling* dengan Metode LDA

Word embedding yang dihasilkan dari tahap sebelumnya digunakan sebagai input dalam pemodelan topik menggunakan metode LDA. Model topik LDA menghitung probabilitas distribusi bersyarat (distribusi posterior) dari variabel tersembunyi berdasarkan variabel yang diamati dengan menggunakan distribusi gabungan. Variabel yang diamati adalah kumpulan kata dan variabel tersembunyi adalah topik [22]. Model representasi grafis LDA ditunjukkan pada Gambar 2.4.



Gambar 2. Model Representasi Grafis LDA [23]

Melalui gambar ini, Faizah dan Lin menjelaskan secara rinci tentang konsep LDA. Pada LDA terdapat beberapa variabel dan parameter yang digunakan. *K* merupakan jumlah topik, *D* adalah jumlah dokumen, dan *N* adalah total jumlah kata dalam seluruh dokumen. Setiap dokumen memiliki distribusi probabilitas atas *K* topik, yang dinotasikan sebagai θ. Secara spesifik, θ_d merupakan distribusi topik untuk dokumen ke-*d*. Selain itu, setiap topik memiliki distribusi probabilitas atas seluruh kata dalam kosakata (*vocabulary*) *V*. Setiap topik dicirikan oleh distribusi atas kata-kata dan dinotasikan sebagai φ_z [23].

Distribusi topik untuk setiap dokumen, θ, dan distribusi kata untuk setiap topik, φ, dapat dimodelkan menggunakan distribusi *Dirichlet* dengan orde *K* dan *V*. *Hyperparameter* dari *prior Dirichlet* untuk distribusi topik per-dokumen dan distribusi kata per-topik, didefinisikan masing-masing sebagai α dan β. Kemudian proses generatif LDA dapat dijelaskan sebagai berikut: (1) pilih distribusi *multinomial* φ_z untuk setiap topik dari distribusi Dirichlet dengan *hyperparameter* β, di mana *z* adalah indeks topik, (2) pilih distribusi *multinomial* θ_d untuk setiap dokumen dari distribusi Dirichlet dengan parameter α, di mana *d* adalah indeks dokumen, (3) pilih topik *z* dari distribusi *multinomial* dengan *hyperparameter* θ_d, (4) pilih kata *w* dari distribusi *multinomial* dengan *hyperparameter* φ_z [23].

Berdasarkan proses di atas, maka probabilitas dari dataset yang diberikan dapat dirumuskan dengan Persamaan 1.

$$p(d,w) = p(d)\sum_z p(z|d)p(w|z) = p(d)\sum_z \theta_d(z)\varphi_z(w) \quad (1)$$

Dimana $p(d,w)$ adalah probabilitas LDA dari *dataset* yang diberikan, yang merupakan gabungan dari probabilitas dokumen *d* dan kata *w* dan $p(d)$ probabilitas dari dokumen *d*. Variabel *z* merujuk pada indeks topik yang dipilih dari distribusi multinomial untuk dokumen tertentu. $p(z|d)$ adalah probabilitas topik *z* diberikan dokumen *d*, yang sama dengan distribusi topik $\theta_d(z)$, sedangkan $p(w|z)$ merupakan probabilitas kata *w* diberikan topik *z*, yang sama dengan distribusi kata $\phi_z(w)$ [23].

2.2. *Topic Modeling* dengan Metode BERTopic

BERTopic, diperkenalkan pada tahun 2020, adalah model topik yang menggunakan transformer BERT, teknik *clustering*, dan variasi c-TF-IDF (*class-based* TF-IDF) untuk menghasilkan representasi topik yang koheren. Model ini dirancang untuk mengatasi keterbatasan model lain yang kurang memperhitungkan hubungan semantik antara kata. BERTopic unggul dalam berbagai *benchmark* berkat kinerjanya yang dapat diskalakan seiring dengan perkembangan model bahasa baru, fleksibilitasnya dalam memisahkan proses *clustering* dari representasi topik, serta kemampuannya untuk memodelkan evolusi topik melalui distribusi kata [24], [25].

Metode BERTopic terdiri dari beberapa langkah utama, yaitu *document embedding, dimensionality reduction, clustering*, dan *topic representation* [10], [24], [26]. Pertama, pada tahap *document embedding*, dokumen dikonversi menjadi representasi numerik berupa vektor berdimensi tinggi yang mewakili isi dan konteks semantik dari dokumen, menggunakan model *sentence-transformers* seperti SBERT yang mengoptimalkan kemiripan semantik. Setelah itu, dilakukan *dimensionality reduction* menggunakan *Uniform Manifold Approximation and Projection* (UMAP) atau metode lain seperti *Principal Component Analysis* (PCA) dan *Truncated Singular Value Decomposition* (TruncatedSVD) untuk mereduksi dimensi vektor dan menjaga struktur data.

Langkah selanjutnya adalah *clustering*, di mana dokumen dikelompokkan berdasarkan kemiripannya menggunakan metode seperti *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) yang dapat mengidentifikasi *outlier* atau metode K-Means, yang membagi data ke dalam k *cluster* berdasarkan jarak terpendek ke *centroid* yang dihitung. Terakhir, dalam tahap *topic representation*, dokumen dalam satu *cluster* digabungkan, frekuensi katanya dihitung menggunakan CountVectorizer, kemudian bobot atau kepentingan kata dalam *cluster* dihitung menggunakan c-TF-IDF untuk menemukan kata kunci yang paling mewakili topik. Kata kunci ini dapat dioptimalkan dengan KeyBERT dan *Maximum Marginal Relevance* (MMR) untuk menghasilkan representasi topik yang lebih relevan dan bervariasi [26].

Proses pemodelan dengan BERTopic melibatkan empat langkah modular, seperti yang telah dijelaskan sebelumnya. Pada *library* Python bertopic, tersedia berbagai algoritma yang dapat dipilih untuk setiap modul tersebut. Peneliti melakukan eksperimen dengan berbagai kombinasi algoritma untuk menemukan hasil yang optimal, menghasilkan model akhir yang mengidentifikasi topik-topik penelitian, daftar kata-kata representatif untuk setiap topik, serta pengelompokan dokumen ke dalam topik-topik tersebut.

2.3. *Evaluation*

Evaluasi dilakukan dengan menghitung *coherence score* dari masing-masing kelompok topik yang dihasilkan. *Coherence score* menilai korelasi semantik antara kata-kata representatif dalam suatu topik, dengan nilai yang lebih tinggi menunjukkan keterkaitan yang lebih kuat dan topik yang lebih mudah diinterpretasikan [14], [19], [27]. *Coherence score* Cv adalah metrik yang sering digunakan untuk mengevaluasi model topik dan menentukan jumlah topik terbaik, sering kali memberikan hasil yang lebih baik dibandingkan ukuran koherensi lainnya [28].

Pada tahapan evaluasi ini ditentukan kelompok topik yang memiliki *coherence score* tertinggi yang dianggap sebagai kelompok topik optimal dan dijadikan dasar untuk tahapan analisis selanjutnya. Evaluasi ini dilakukan pada model-model dari kedua metode, baik LDA maupun BERTopic, dengan metode eksperimen berdasarkan parameter model, seperti jumlah topik (*num_topics*) yang dihasilkan atau ukuran minimal *cluster* (*min_cluster_size*), untuk menghasilkan model terbaik dari masing-masing metode. Model-model terbaik ini selanjutnya digunakan untuk menganalisis tren penelitian di bidang Ilmu Komputer selama lima tahun terakhir (2019-2023). Kedua model ini juga dibandingkan dan dianalisis untuk menentukan metode terbaik pada kasus penelitian ini.

2.4. *Topic Trend Analysis*

Topik optimal yang dihasilkan dari analisis sebelumnya divisualisasikan untuk mengamati tren penelitian dengan menggunakan diagram garis (*line chart*) yang menggambarkan frekuensi publikasi setiap topik penelitian per tahun. Visualisasi ini mencerminkan dinamika perkembangan topik berdasarkan jumlah publikasi per tahun. Selain itu, *heatmap* juga digunakan untuk memperlihatkan fluktuasi tren penelitian melalui gradasi warna yang menunjukkan perubahan indeks publikasi tahunan dari 2019 hingga 2023. Indeks +1 menunjukkan peningkatan jumlah publikasi sebesar 100% dibandingkan tahun sebelumnya, sedangkan indeks -1 menunjukkan penurunan 100%. Melalui visualisasi ini, tren penelitian dari tahun ke tahun dapat diidentifikasi dengan jelas.

3. Hasil dan Pembahasan

Data penelitian awalnya berjumlah 6.000, namun setelah melalui tahapan persiapan data, hanya tersisa 4.892 data. Dari lima fitur yang ada, dua fitur yang digunakan untuk *topic modeling* dan analisis tren penelitian adalah tahun publikasi (*Year*) dan data teks (*Text*). Fitur-fitur ini kemudian diproses dalam tahapan analisis lebih lanjut.

3.1. Hasil *Topic Modeling* dengan LDA

Metode LDA dalam penelitian ini diterapkan dengan *embedding* BoW dan TF-IDF yang dibentuk dari fitur *Text* setelah melalui proses *text preprocessing*. *Topic modeling* dengan LDA didasarkan pada teori dan

Persamaan (1) yang telah diuraikan sebelumnya, diimplementasikan dengan *library* Python gensim. Hasilnya berupa daftar topik yang ditemukan dan kata-kata representatifnya, serta hasil pengelompokan dokumen ke dalam topik-topik tersebut. Berikut adalah hasil penerapan *topic modeling* dengan LDA menggunakan *embedding* BoW, disajikan dalam Tabel 1 dan 2.

Tabel 1.Hasil *Topic Modeling* dengan LDA: Daftar Topik

ID Topik	Kata-Kata Representatif dan Probabilitasnya
0	0.030*"text" + 0.024*"mining" + 0.017*"review" + 0.016*"data" + 0.015*"sentiment" + 0.014*"analysis" + 0.014*"use" + 0.011*"quality" + 0.011*"tourism" + 0.010*"approach"
1	0.022*"model" + 0.014*"control" + 0.013*"use" + 0.009*"method" + 0.007*"result" + 0.007*"nonlinear" + 0.006*"system" + 0.006*"study" + 0.006*"paper" + 0.006*"energy"
2	0.032*"data" + 0.013*"study" + 0.011*"big" + 0.011*"firm" + 0.010*"use" + 0.010*"ai" + 0.008*"research" + 0.008*"paper" + 0.008*"financial" + 0.007*"author"
3	0.022*"research" + 0.018*"study" + 0.015*"paper" + 0.012*"review" + 0.011*"literature" + 0.011*"analysis" + 0.010*"industry" + 0.009*"technology" + 0.009*"use" + 0.009*"business"
4	0.025*"study" + 0.021*"project" + 0.015*"management" + 0.013*"performance" + 0.012*"use" + 0.009*"model" + 0.008*"relationship" + 0.008*"knowledge" + 0.008*"process" + 0.008*"research"
5	0.017*"study" + 0.012*"use" + 0.008*"social" + 0.006*"covid" + 0.006*"service" + 0.006*"paper" + 0.006*"data" + 0.005*"health" + 0.005*"information" + 0.005*"research"
6	0.019*"model" + 0.019*"use" + 0.016*"prediction" + 0.010*"neural" + 0.009*"train" + 0.009*"flight" + 0.009*"predict" + 0.009*"signal" + 0.009*"aircraft" + 0.008*"accuracy"
7	0.021*"user" + 0.020*"use" + 0.020*"study" + 0.018*"consumer" + 0.015*"customer" + 0.014*"social" + 0.012*"service" + 0.012*"online" + 0.011*"intention" + 0.010*"model"
8	0.039*"blockchain" + 0.028*"technology" + 0.025*"chain" + 0.025*"supply" + 0.013*"study" + 0.012*"information" + 0.011*"research" + 0.011*"paper" + 0.010*"framework" + 0.008*"adoption"
9	0.017*"propose" + 0.015*"use" + 0.015*"algorithm" + 0.014*"data" + 0.013*"network" + 0.009*"paper" + 0.009*"method" + 0.006*"base" + 0.006*"image" + 0.006*"application"
10	0.029*"system" + 0.014*"use" + 0.014*"model" + 0.012*"design" + 0.012*"propose" + 0.012*"method" + 0.009*"paper" + 0.008*"fuzzy" + 0.008*"control" + 0.007*"base"

Tabel 1 menunjukkan daftar topik teridentifikasi dari proses *topic modeling* dengan LDA menggunakan representasi BoW, dengan batasan jumlah topik yang ditentukan yaitu *num_topics*=11. Untuk merepresentasikan topik-topik tersebut, diidentifikasi 10 kata dengan probabilitas tertinggi dalam masing-masing topik. Misalnya, Topik 0 berkaitan dengan analisis teks atau *text mining*, khususnya dalam konteks analisis sentimen dan kualitas data, dengan aplikasi yang mungkin mencakup sektor pariwisata. Kata-kata representatif seperti "*text*," "*mining*," "*sentiment*," dan "*tourism*" menunjukkan fokus pada pemrosesan data tekstual untuk mengevaluasi kualitas atau opini, contohnya dalam industri pariwisata.

Seluruh dokumen dalam *dataset* selanjutnya dikelompokkan ke salah satu dari 10 topik tersebut. Cuplikan hasilnya disajikan pada Tabel 2 berikut.

Tabel 2.Hasil *Topic Modeling* dengan LDA: Pengelompokan Dokumen

No.	DOI	Judul Artikel (*Title*)	Topik	Probabilitas Topik
1	10.1108/IJWIS-02-2022-0044	*Fake news detection on Twitter*	0	{0: 0.3775177, 5: 0.061035596, 6: 0.08950299, 7: 0.26739565, 9: 0.20223516}
2	10.1108/IMDS-05-2021-0323	*A progressive genetic-based neural architecture search*	9	{6: 0.09422406, 7: 0.10505944, 9: 0.5310757, 10: .26660153}
3	10.1108/IJICC-06-2021-0109	*High accuracy offering attention mechanisms based deep learning approach using CNN/bi-LSTM for sentiment analysis*	9	{0: 0.1898691, 6: 0.34752366, 9: 0.45709258}
4	10.1108/WJE-04-2021-0204	*An intrusion detection system for health-care system using machine and deep learning*	9	{3: 0.12529536, 6: 0.15737039, 9: 0.609753, 10: 0.10252378}
5	10.1108/COMPEL-12-2022-0436	*Evaluating magnetic fields using deep learning*	10	{1: 0.0899217, 6: 0.13861789, 8: 0.066461466, 9: 0.3489543, 10: 0.35377708}
…	…	…	…	…
4892	10.24006/jilt.2021.19.2.096	*Exploring the Characteristics of High-Speed Rail and Air Transportation Networks in China: A Weighted Network Approach*	10	{0: 0.16334012, 3: 0.14348388, 6: 0.18647473, 7: 0.13973774, 9: 0.0968097, 10: 0.2674275}

Tabel 2 menyajikan hasil pengelompokan dokumen dalam *dataset* berdasarkan topik yang paling relevan. Mengacu pada konsep LDA, yang mengasumsikan bahwa setiap dokumen terdiri dari beberapa topik, terlihat bahwa sejumlah dokumen dapat memiliki lebih dari satu topik. Contohnya, dokumen ke-1 memiliki 5 topik, yaitu topik 0, 5, 6, 7, dan 9. Namun, dalam pengelompokan akhir, dokumen ini dimasukkan ke dalam kelompok Topik 0 karena memiliki probabilitas topik-dokumen tertinggi, yaitu 0,3775177. Penelitian dalam dokumen tersebut membahas deteksi berita palsu di media sosial Twitter dengan menggunakan data teks dari berbagai posting di platform tersebut. Hal ini sesuai dengan pengelompokannya dalam Topik 0, yang

berfokus pada analisis data teks. Dengan demikian, penerapan *topic modeling* menggunakan LDA menghasilkan pengelompokan dokumen berdasarkan probabilitas topik yang ada di dalam dokumen tersebut.

3.2. Hasil *Topic Modeling* dengan BERTopic

Metode BERTopic yang digunakan dalam penelitian ini dibangun tanpa melalui tahap *text preprocessing*. BERTopic dapat dikonfigurasi dengan berbagai kombinasi modul, salah satunya adalah yang menggunakan metode TruncatedSVD untuk *dimensionality reduction* dan K-Means untuk *clustering*. Hasil penerapan BERTopic dengan TruncatedSVD-KMeans ini disajikan pada Tabel 3 dan Tabel 4.

Tabel 3. Hasil *Topic Modeling* dengan BERTopic: Daftar Topik

ID Topik	Kata-Kata Representatif dan Nilai c-TF-IDF-nya
0	0.424 * "projects" + 0.412 * "management" + 0.397 * "project" + 0.361 * "organisational" + 0.348 * "leadership" + 0.323 * "stakeholder" + 0.323 * "stakeholders" + 0.322 * "organizational" + 0.317 * "managers" + 0.317 * "organization"
1	0.413 * "business" + 0.385 * "marketing" + 0.375 * "analytics" + 0.368 * "management" + 0.366 * "innovation" + 0.363 * "companies" + 0.336 * "industry" + 0.334 * "data" + 0.333 * "firms" + 0.327 * "technology"
2	0.438 * "ethical" + 0.429 * "ethics" + 0.367 * "companies" + 0.360 * "csr" + 0.345 * "corporate" + 0.338 * "accountability" + 0.338 * "sustainability" + 0.329 * "organizations" + 0.327 * "accounting" + 0.307 * "governance"
3	0.240 * "motor" + 0.222 * "controller" + 0.202 * "control" + 0.143 * "nonlinear" + 0.140 * "electric" + 0.112 * "robot" + 0.097 * "sensor" + 0.069 * "motion" + 0.055 * "tracking" + 0.049 * "aircraft"
4	0.448 * "marketing" + 0.405 * "social" + 0.370 * "media" + 0.366 * "consumers" + 0.357 * "customers" + 0.344 * "twitter" + 0.339 * "privacy" + 0.334 * "technology" + 0.333 * "consumer" + 0.332 * "communication"
5	0.304 * "reliability" + 0.280 * "maintenance" + 0.262 * "optimization" + 0.242 * "manufacturing" + 0.237 * "assembly" + 0.234 * "robot" + 0.228 * "industrial" + 0.214 * "parts" + 0.197 * "components" + 0.186 * "algorithm"
6	0.398 * "vr" + 0.393 * "customers" + 0.381 * "immersive" + 0.376 * "marketing" + 0.363 * "consumers" + 0.352 * "retail" + 0.343 * "products" + 0.341 * "customer" + 0.337 * "consumer" + 0.324 * "shopping"
7	0.540 * "blockchain" + 0.292 * "stakeholders" + 0.290 * "governance" + 0.283 * "business" + 0.262 * "organizations" + 0.253 * "industry" + 0.252 * "management" + 0.248 * "companies" + 0.243 * "chain" + 0.242 * "security"
8	0.476 * "lstm" + 0.439 * "predicting" + 0.380 * "classifier" + 0.364 * "forecasting" + 0.353 * "convolutional" + 0.344 * "prediction" + 0.315 * "predict" + 0.299 * "classification" + 0.279 * "datasets" + 0.274 * "neural"
9	0.512 * "analytics" + 0.492 * "data" + 0.351 * "accounting" + 0.321 * "insights" + 0.319 * "analysis" + 0.304 * "management" + 0.300 * "business" + 0.296 * "information" + 0.272 * "investors" + 0.271 * "ai"
10	0.398 * "iot" + 0.349 * "logistics" + 0.313 * "blockchain" + 0.300 * "industry" + 0.289 * "business" + 0.284 * "management" + 0.265 * "analytics" + 0.263 * "technologies" + 0.254 * "technology" + 0.250 * "industrial"
11	0.346 * "management" + 0.342 * "projects" + 0.320 * "construction" + 0.319 * "project" + 0.315 * "industry" + 0.305 * "industries" + 0.299 * "business" + 0.278 * "manufacturing" + 0.273 * "organization" + 0.271 * "design"
12	0.442 * "iot" + 0.347 * "routing" + 0.283 * "sensors" + 0.269 * "sensor" + 0.257 * "wireless" + 0.244 * "encryption" + 0.244 * "algorithm" + 0.242 * "algorithms" + 0.234 * "node" + 0.231 * "security"

Tabel 3 menyajikan topik yang dihasilkan dari pemodelan topik menggunakan BERTopic dengan SBERT sebagai embedding, TruncatedSVD untuk reduksi dimensi, dan K-Means untuk *clustering*, dengan jumlah topik yang ditentukan sebanyak 13 (*num_topics*=13). Setiap topik direpresentasikan oleh 10 kata paling representatif. Misalnya, Topik 0 berfokus pada aspek manajemen proyek dalam organisasi. Kata-kata representatif seperti "*projects*," "*management*," "*project*," "*organisational*," dan "*leadership*" menunjukkan keterkaitan topik dengan manajemen proyek dalam organisasi. Selain itu, istilah seperti "*stakeholder*," "*managers*," dan "*organization*" mengindikasikan perhatian terhadap peran pemangku kepentingan dan manajer dalam kesuksesan proyek. Dalam konteks penelitian bidang ilmu komputer, topik ini mungkin terkait dengan manajemen proyek perangkat lunak, pengembangan sistem informasi, atau studi tentang peran organisasi dalam implementasi teknologi.

Beberapa topik menunjukkan kesamaan dalam kata-kata representatif, seperti pada Topik 10 dan Topik 12, di mana keduanya memiliki kata "*iot*" sebagai kata representatif pertama. Meskipun begitu, analisis lebih lanjut menunjukkan perbedaan fokus antara kedua topik tersebut. Topik 10 berpusat pada aplikasi *Internet of Things* (IoT) dalam industri dan logistik, dengan kata-kata seperti "*logistics*," "*blockchain*," dan "*industry*" yang menunjukkan penggunaan IoT untuk efisiensi bisnis. Sementara itu, Topik 12 lebih menekankan aspek teknis dan keamanan IoT, dengan kata-kata seperti "*routing*," "*sensors*," dan "*security*" yang mencerminkan fokus pada algoritma routing dan isu-isu enkripsi. Perbedaan utama terletak pada fokus praktis di Topik 10 dan fokus teknis di Topik 12.

Model topik ini kemudian digunakan pada pengelompokan dokumen. Seluruh dokumen dalam data dikelompokkan ke salah satu dari 13 topik yang teridentifikasi. Cuplikan hasilnya dapat dilihat pada Tabel 4.

Tabel 4. Hasil *Topic Modeling* dengan BERTopic: Pengelompokan Dokumen

No.	DOI	Judul Artikel (*Title*)	Topik

1	10.1108/IJWIS-02-2022-0044	*Fake news detection on Twitter*	8
2	10.1108/IMDS-05-2021-0323	*A progressive genetic-based neural architecture search*	8
3	10.1108/IJICC-06-2021-0109	*High accuracy offering attention mechanisms based deep learning approach using CNN/bi-LSTM for sentiment analysis*	8
4	10.1108/WJE-04-2021-0204	*An intrusion detection system for health-care system using machine and deep learning*	8
5	10.1108/COMPEL-12-2022-0436	*Evaluating magnetic fields using deep learning*	8
…	…	…	…
4892	10.24006/jilt.2021.19.2.096	*Exploring the Characteristics of High-Speed Rail and Air Transportation Networks in China: A Weighted Network Approach*	9

Tabel 4 menunjukkan hasil pengelompokan dokumen dalam dataset berdasarkan topik yang paling relevan, misalnya dokumen ke-1 dikelompokkan ke Topik 8 yang berfokus pada penerapan model deep learning, khususnya LSTM, untuk prediksi dan klasifikasi. Kata-kata kunci seperti "lstm," "*predicting*," dan "*classifier*" (Tabel 3) mengindikasikan bahwa topik ini terkait dengan penggunaan teknik pembelajaran mesin untuk berbagai aplikasi seperti prediksi cuaca, analisis sentimen, atau pengenalan pola. Pengelompokan serupa juga dilakukan untuk dokumen lainnya, sesuai dengan isi dan fokus penelitian masing-masing.

3.3. Evaluasi Hasil

Setelah melakukan eksplorasi dengan berbagai teknik pada kedua metode *topic modeling*, yaitu LDA dan BERTopic, semua model yang dihasilkan kemudian dievaluasi. Evaluasi ini bertujuan untuk menentukan model yang paling unggul berdasarkan metrik *coherence score*. Hasil perhitungan *coherence score* tersebut disajikan dalam Tabel 5.

Tabel 5. Hasil Evaluasi Model Topik

No.	Model Topik	*Coherence Score*	*Outlier*
1	LDA-BoW (num_topics=11)	0.4223	0
2	LDA-TFIDF (num_topics=17)	0.6868	0
3	BERTopic (UMAP-HDBSCAN; min_cluster_size =95)	0.4578	1403
4	BERTopic (PCA-HDBSCAN; min_cluster_size =70)	0.5221	3824
5	BERTopic (TruncatedSVD-HDBSCAN; min_cluster_size=50)	0.5351	3972
6	BERTopic (UMAP-KMeans; num_topics=16)	0.4384	0
7	BERTopic (PCA-KMeans; num_topics=16)	0.4594	0
8	BERTopic (TruncatedSVD-KMeans; num_topics=13)	0.4898	0

Tabel 5 menunjukkan bahwa model LDA dengan *feature extraction* menggunakan TF-IDF mencatatkan *coherence score* tertinggi, yaitu 0,6868 (≈ 0,69). Selanjutnya, model BERTopic TruncatedSVD-HDBSCAN mencapai *coherence score* sebesar 0,5351 (≈ 0,54). Sebaliknya, model LDA yang menggunakan feature extraction dengan BoW memiliki *coherence score* terendah, yaitu 0,4223 (≈ 0,49).

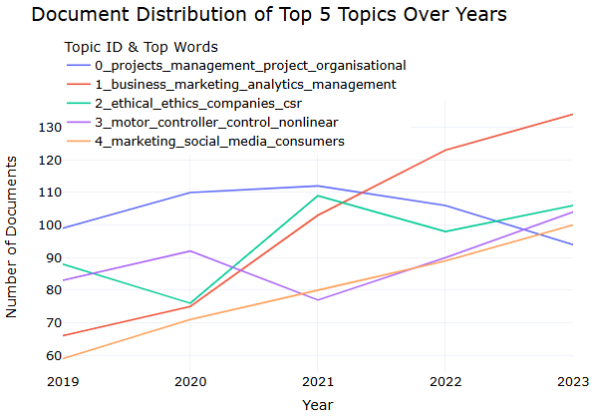
Evaluasi lebih lanjut terhadap distribusi dokumen hasil pengelompokan mengungkapkan kelemahan signifikan pada model LDA-TFIDF. Mayoritas dokumen terkonsentrasi dalam satu topik tunggal (Topik 15 dengan 4.878 dokumen), menunjukkan bahwa model ini tidak efektif dalam mengidentifikasi variasi topik dalam *dataset*. Di sisi lain, model LDA-BoW, meskipun memiliki *coherence score* yang lebih rendah, mampu menghasilkan distribusi dokumen yang lebih merata, dengan jumlah dokumen per topik berkisar antara 792–39 dokumen. Karakteristik ini menjadikan model LDA-BoW lebih interpretatif dan efektif untuk analisis lanjutan.

Analisis lebih lanjut terhadap jumlah *outlier* yang teridentifikasi mengungkapkan keterbatasan signifikan pada model BERTopic dengan metode HDBSCAN. Model TruncatedSVD-HDBSCAN, meskipun memiliki *coherence score* tinggi, mengidentifikasi sejumlah besar data sebagai *outlier*—mencapai 81,19% dari total data. Fenomena ini mengakibatkan pengurangan substansial dalam jumlah data yang dapat dianalisis, mengindikasikan ketidaksesuaian metode HDBSCAN untuk *dataset* berskala kecil. Berdasarkan pertimbangan ini, model BERTopic dengan metode *clustering* K-Means, khususnya model TruncatedSVD-KMeans dengan *coherence score* 0,4898 (≈ 0,49), dipilih sebagai model optimal untuk analisis selanjutnya.

Pemilihan model TruncatedSVD-KMeans didasarkan pada *coherence score* yang lebih tinggi dibandingkan dengan model LDA yang sebelumnya dipertimbangkan. Model ini selanjutnya digunakan dalam analisis tren penelitian di bidang Ilmu Komputer. Secara keseluruhan, metode *topic modeling* BERTopic menunjukkan efektivitas yang lebih tinggi dalam menganalisis topik penelitian dibandingkan dengan LDA, terutama dalam aspek *coherence score* yang dihasilkan. Temuan ini sejalan dengan hasil penelitian-penelitian terdahulu yang menyimpulkan bahwa metode BERTopic lebih efektif dan dapat menghasilkan topik-topik yang lebih koheren [18], [19].

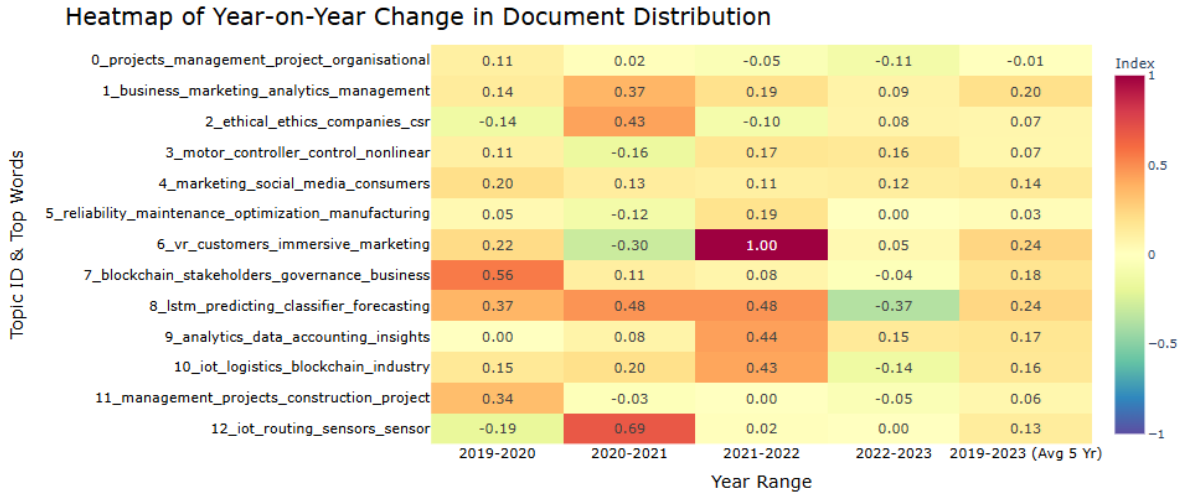
3.4 Analisis Tren Penelitian

Penelitian di bidang Ilmu Komputer dari tahun ke tahun terus mengalami peningkatan, namun jika ditelusuri lebih mendalam, peningkatan tersebut bisa jadi hanya terjadi di beberapa subbidang atau topik penelitian, sementara topik lain mengalami penurunan minat seiring munculnya tren dan teknologi baru. Berdasarkan model topik BERTopic terpilih (TruncatedSVD-KMeans), terdapat 13 topik penelitian di bidang Ilmu Komputer. Tren penelitian pada 5 topik dengan jumlah publikasi terbanyak diilustrasikan pada Gambar 3.



Gambar 3. Tren Penelitian 5 Topik dengan Publikasi Terbanyak

Gambar 3 menunjukkan bahwa, dari 5 topik dengan publikasi terbanyak, topik bisnis, pemasaran, dan analitik (Topik 1 dan 4) menunjukkan pertumbuhan konsisten dan signifikan. Sementara itu, topik manajemen proyek (Topik 0) mengalami penurunan gradual, dan topik etika (Topik 2) serta kontrol motor (Topik 3) menunjukkan fluktuasi namun tetap relevan. Temuan ini mengindikasikan pergeseran fokus penelitian ke arah aspek bisnis dan analitik data dalam Ilmu Komputer. Tren secara keseluruhan yang melibatkan semua topik penelitian disajikan dalam bentuk *heatmap* indeks peningkatan jumlah publikasi pada Gambar 3.



Gambar 4. Indeks Peningkatan Publikasi per Topik Penelitian

Berdasarkan Gambar 4 diketahui dinamika tren penelitian dalam bidang Ilmu Komputer dari 2019 hingga 2023 yang menunjukkan pergeseran minat secara signifikan. Topik manajemen proyek (Topik 0) mengalami penurunan tajam sejak 2022, dengan penurunan sebesar 11% pada 2023, mengindikasikan penurunan minat atau kejenuhan. Sebaliknya, topik analitik bisnis dan pemasaran (Topik 1) menunjukkan pertumbuhan konsisten, dengan peningkatan rata-rata tahunan 20% dan puncaknya mencapai 37% pada 2021, mencerminkan permintaan yang meningkat dalam konteks bisnis modern.

Topik teknik prediksi (Topik 8) seperti *deep learning* dengan LSTM dan metode lainnya menunjukkan pertumbuhan kuat hingga 2022, namun mengalami penurunan pada 2023. Topik analitik data dalam akuntansi (Topik 9) juga mengalami peningkatan tajam sebesar 44% pada 2022 setelah periode stabil sebelumnya, menandakan meningkatnya perhatian terhadap analitik dalam bidang ini. Sementara itu, topik terkait teknologi canggih seperti *blockchain* (Topik 7) dan IoT dalam logistik (Topik 10) menunjukkan lonjakan minat pada tahun-tahun tertentu. Fluktuasi juga terlihat pada topik pemasaran VR (Topik 6) dan IoT *routing* (Topik 12), dengan penurunan tajam di satu tahun diikuti lonjakan di tahun berikutnya, mencerminkan adanya ketidakstabilan dalam minat atau fokus penelitian terhadap topik-topik ini.

Secara keseluruhan, tren penelitian menunjukkan pergeseran dari topik tradisional seperti manajemen proyek menuju fokus yang lebih besar pada analitik bisnis, *blockchain*, IoT, dan *deep learning*, mencerminkan kemajuan teknologi dan perubahan kebutuhan industri. Temuan-temuan ini memberikan wawasan berharga tentang arah perkembangan penelitian di bidang Ilmu Komputer. Wawasan tersebut dapat dimanfaatkan oleh peneliti, institusi pendidikan, dan industri untuk mengarahkan fokus penelitian, mengalokasikan sumber daya, serta mengantisipasi kebutuhan dan tren teknologi di masa depan.

4. Kesimpulan

Penerapan metode *topic modeling* BERTopic dan LDA berhasil mengidentifikasi topik-topik utama dalam penelitian Ilmu Komputer. Model LDA dengan *embedding* BoW menghasilkan 11 topik, sementara BERTopic dengan kombinasi TruncatedSVD-KMeans menghasilkan 13 topik. Evaluasi menunjukkan bahwa model BERTopic memiliki *coherence score* lebih tinggi (0,49) dibandingkan LDA-BoW (0,42), mengindikasikan bahwa BERTopic lebih efektif dalam

menghasilkan topik-topik yang koheren dan lebih mudah diinterpretasikan.

Analisis tren penelitian periode 2019-2023 mengungkapkan dinamika yang beragam dalam bidang Ilmu Komputer. Topik-topik seperti analitik bisnis dan pemasaran menunjukkan pertumbuhan yang konsisten, dengan rata-rata peningkatan mencapai 20% selama lima tahun. Teknologi *blockchain* dan IoT juga mencatatkan pertumbuhan yang pesat pada tahun-tahun tertentu. Sebaliknya, topik-topik seperti VR dan teknik prediksi mengalami fluktuasi yang signifikan. Terjadi pergeseran fokus penelitian menuju analitik bisnis, *blockchain*, IoT, dan teknik prediksi seperti *deep learning*, sementara topik-topik tradisional seperti manajemen proyek cenderung mengalami penurunan atau pertumbuhan yang lebih lambat. Temuan ini dapat dimanfaatkan untuk mengarahkan fokus penelitian dan mengantisipasi tren teknologi di masa depan. Untuk penelitian selanjutnya, disarankan untuk meningkatkan jumlah dan cakupan periode data, memanfaatkan data dari berbagai sumber penelitian bereputasi, serta mengeksplorasi metode *topic modeling* lainnya guna meningkatkan efektivitas model dan koherensi topik yang dihasilkan.

Daftar Rujukan

[1] Shu, X., & Ye, Y. (2023). Knowledge discovery: Methods from data mining and machine learning. *Social Science Research*, *110*, 102817. https://doi.org/10.1016/j.ssresearch.2022.102817

[2] Jansevskis, M., & Osis, K. (2023). Knowledge discovery frameworks and characteristics. *Baltic Journal of Modern Computing*, *11*(4), 686–702. https://doi.org/10.22364/bjmc.2023.11.4.08

[3] Głowania, S., Kozak, J., & Juszczuk, P. (2023). Knowledge discovery in databases for a football match result. *Electronics*, *12*(12), 2712. https://doi.org/10.3390/electronics12122712

[4] Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. *Entropy*, *23*(4), 485. https://doi.org/10.3390/e23040485

[5] Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems*, *121*, 102342. https://doi.org/10.1016/j.is.2023.102342

[6] Aleqabie, H. J., Sfoq, M. S., Albeer, R. A., & Abd, E. H. (2024). A review of text mining techniques: Trends, and applications in various domains. *Iraqi Journal for Computer Science and Mathematics*. https://doi.org/10.52866/ijcsm.2024.05.01.009

[7] Birunda, S. S., & Devi, R. K. (2021). A review on word embedding techniques for text classification. In *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020* (pp. 267–281). Springer. https://doi.org/10.1007/978-981-15-9651-3_23

[8] Khairunnisa, S., Adiwijaya, A., & Faraby, S. A. (2021). Pengaruh text preprocessing terhadap analisis sentimen komentar masyarakat pada media sosial Twitter (Studi kasus pandemi COVID-19). *Jurnal Media Informatika Budidarma*, *5*(2), 406. https://doi.org/10.30865/mib.v5i2.2835

[9] de Lima, B. C., Baracho, R. M. A., Mandl, T., & Porto, P. B. (2023). Reactions to science communication: Discovering social network topics using word embeddings and semantic knowledge. *Social Network Analysis and Mining*, *13*(1). https://doi.org/10.1007/s13278-023-01125-5

[10] Niroomand, K., Saady, N. M. C., Bazan, C., Zendehboudi, S., Soares, A., & Albayati, T. M. (2023). Smart investigation of artificial intelligence in renewable energy system technologies by natural language processing: Insightful pattern for decision-makers. *Engineering Applications of Artificial Intelligence*, *126*, 106848. https://doi.org/10.1016/j.engappai.2023.106848

[11] Takacs, V., & O'Brien, C. D. (2023). Trends and gaps in biodiversity and ecosystem services research: A text mining approach. *Ambio*, *52*(1), 81–94. https://doi.org/10.1007/s13280-022-01776-2

[12] Scopus. (n.d.). Scopus. *Elsevier*. https://www.scopus.com/ (accessed 21 August 2024).

[13] Qin, H., Zeng, J., & Ma, X. (2021). Trend analysis of research direction in computer science based on Microsoft academic graph. In *Proceedings of ACM International Conference* (Vol. Part F168982). Association for Computing Machinery. https://doi.org/10.1145/3448734.3450470

[14] Samsir, Saragih, R. S., Subagio, S., Aditya, R., & Watrianthos, R. (2023). BERTopic modeling of natural language processing abstracts: Thematic structure and trajectory. *Jurnal Media Informatika Budidarma*, *7*(3), 1514–1520. https://doi.org/10.30865/mib.v7i3.6426

[15] Akbarighatar, P., Pappas, I., & Vassilakopoulou, P. (2023). A sociotechnical perspective for responsible AI maturity models: Findings from a mixed-method literature review. *International Journal of Information Management Data Insights*, *3*(2), 100193. https://doi.org/10.1016/j.jjimei.2023.100193

[16] Yu, D., Fang, A., & Xu, Z. (2023). Topic research in fuzzy domain: Based on LDA topic modelling. *Information Sciences*, *648*, 119600. https://doi.org/10.1016/j.ins.2023.119600

[17] Jung, Y. J., & Kim, Y. (2023). Research trends of sustainability and marketing research, 2010–2020: Topic modeling analysis. *Heliyon*, *9*(3), e14208. https://doi.org/10.1016/j.heliyon.2023.e14208

[18] Kukushkin, K., Ryabov, Y., & Borovkov, A. (2022). Digital twins: A systematic literature review based on data analysis and topic modeling. *Data*, *7*(12), 173. https://doi.org/10.3390/data7120173

[19] Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of topic modelling approaches in the banking context. *Applied Sciences*, *13*(2), 797. https://doi.org/10.3390/app13020797

[20] Widyaningsih, T. W., Dewi, M. A., & Andrianingsih, A. (2021). Analisis bibliometrik untuk memetakan tren penelitian COVID-19 dalam topik ilmu komputer. *Techno. Com*, *20*(3), 440–454. https://doi.org/10.33633/tc.v20i3.4593

[21] Kumar, A. N., Raj, R. K., Aly, S. G., Anderson, M. D., Becker, B. A., Blumenthal, R. L., Eaton, E., et al. (2024). *Computer Science Curricula 2023*. Association for Computing Machinery. https://doi.org/10.1145/3664191

[22] Gan, J., & Qi, Y. (2021). Selection of the optimal number of topics for LDA topic model—Taking patent policy analysis as an example. *Entropy*, *23*(10), 1301. https://doi.org/10.3390/e23101301

[23] Faizah, & Lin, B. S. (2023). Visualizing change and correlation of topics with LDA and agglomerative clustering on COVID-19 vaccine tweets. *IEEE Access*, *11*, 51647–51656. https://doi.org/10.1109/ACCESS.2023.3278979

[24] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://doi.org/10.48550/arXiv.2203.05794

[25] Mendonça, M., & Figueira, Á. (2024). Topic extraction: BERTopic's insight into the 117th Congress's Twitterverse. *Informatics*, *11*(1), 8. https://doi.org/10.3390/informatics11010008

[26] Lu, C., Zhu, L., Xie, Y., Xu, W., Zhao, Y., & Cao, Y. (2024). Analysis of hot topics and evolution of research in world-class agricultural universities based on BERTopic. *Applied Mathematics and Nonlinear Sciences*, *9*(1). https://doi.org/10.2478/amns-2024-0327

[27] Khadija, M. A., & Nurharjadmo, W. (2024). Enhancing Indonesian customer complaint analysis: LDA topic modelling

with BERT embeddings. *Sinergi, 28*(1), 153–162. https://doi.org/10.22441/sinergi.2024.1.015

[28] Hananto, V.R. (2023). Implementation of Dynamic Topic Modeling to Discover Topic Evolution on Customer Reviews. *Jurnal Online Informatika, 8*(2), 147–157, https://doi.org/10.15575/join.v8i2.963