Indonesian Journal of Multidisciplinary on Social and Technology
Homepage: https://journal.ilmudata.co.id/index.php/ijmst
Vol. 4 No. 3 (2026) pp: 10895-10914
P-ISSN: 2986-6790, e-ISSN: 2986-6782

A Comparative Analysis of LDA and BERTopic for Identifying Research Topic Trends in Computer Science Using Journal Article Titles

M. Zam Zami, Andreo Yudertha
Information Systems Study Program, Faculty of Science and Technology, Universitas Islam Negeri Sulthan Thaha Saifuddin Jambi, Jambi, Indonesia.
zamzamiii155@gmail.com, andreo@uinjambi.ac.id*

Abstract

Topic modeling from journal article titles is challenging because titles are short and provide limited word co occurrence and contextual information. This study compares Latent Dirichlet Allocation (LDA) and BERTopic for identifying computer science research topics and examining their temporal trends. The corpus comprises 39,532 titles published between 2015 and 2025 across 283 Indonesian journals indexed in the Science and Technology Index (SINTA). Of these, 46.2% were originally written in English and translated into Indonesian before modeling. Both approaches were evaluated using the same corpus, vocabulary, topic coherence (C_v), topic diversity, and three random seeds. LDA was tested across topic counts from 5 to 30, while BERTopic used multilingual sentence embeddings, UMAP, and HDBSCAN and was reduced to 25 topics for comparison. Temporal trends for 2015–2024 were assessed using three statistical tests and robustness checks for journal composition, SINTA ranking, and original title language. At 25 topics, BERTopic achieved higher coherence (0.671, SD = 0.017) and diversity (0.841) than LDA (0.580, SD = 0.039; diversity = 0.693), but left 37.9% of titles unassigned. LDA assigned all titles and produced a more balanced topic-size distribution. Agreement between the partitions was limited (NMI = 0.31), with 11 of 25 LDA topics having one-to-one counterparts. Seven method–topic combinations passed all reported trend checks, including sentiment analysis, deep learning, and IoT monitoring, whereas some steep declines reflected changes in journal composition. The findings reveal a trade-off between topic quality and document coverage and indicate that research trends should be interpreted only when they remain robust across methods and corpus-composition checks.

Keywords: Topic Modeling, Latent Dirichlet Allocation, BERTopic, Research Trends, Short Texts, Computer Science, Bibliometrics, Indonesia

1. Introduction

Understanding how research topics emerge, persist, or decline is important for research assessment, science policy, and research agenda planning. Topic models make it possible to summarize large collections of scholarly documents into interpretable themes. However, article titles are short and often contain limited context, which makes topic identification more difficult than modeling abstracts or full texts.

Latent Dirichlet Allocation (LDA) represents each document as a mixture of topics over a bag-of-words vocabulary [3]. BERTopic, in contrast, embeds documents using a sentence-embedding model and clusters semantically similar representations before describing each cluster using class-based term frequency–inverse document frequency (c-TF-IDF) [4]. Previous comparative research has evaluated LDA, NMF, Top2Vec, and BERTopic on social media data, highlighting the value of comparing topic modeling methods across different textual contexts [5]. Related research has also explored the integration of word embeddings into probabilistic topic models. Dieng et al. [6], for example, introduced the Embedded Topic Model, which combines word embeddings with a probabilistic topic modeling framework.These approaches therefore capture different signals: LDA relies on lexical co-occurrence, whereas BERTopic relies on contextual embeddings and density based clustering.

The corpus contains 39,532 article titles published between 2015 and 2025 in 283 Indonesian journals indexed by SINTA (Science and Technology Index). Three characteristics make the comparison challenging: the titles are short, 46.2% were originally written in English and translated into Indonesian, and the set of contributing journals changes substantially over time. These characteristics may affect both topic quality and apparent temporal trends. This study addresses four research questions:

RQ1. Which research topics can be identified from computer science journal article titles published during 2015–2025?

RQ2. How do LDA and BERTopic differ in identifying topics from short journal article titles, including their topic assignments and document coverage?

RQ3. How do LDA and BERTopic compare in topic coherence, topic diversity, interpretability, and suitability for identifying research trends from short academic titles?

RQ4. How did the identified research topics evolve from 2015 to 2025, and which temporal patterns remain robust under statistical and corpus-composition checks?

The study makes three contributions. Methodologically, it compares LDA and BERTopic on short academic titles using a shared corpus and evaluation protocol. Empirically, it reports topic quality alongside assignment coverage and cross-model alignment. For trend analysis, it evaluates candidate temporal patterns with statistical tests and robustness checks for changes in journal composition, SINTA ranking, and original title language.

2. Related Work

Topic models for scholarly literature. Probabilistic topic models have long been used to discover themes in scientific corpora [1]. LDA [3] remains a standard approach; its results are often interpretable but may be affected by sparse co-occurrence in short texts. Specialized approaches such as the biterm topic model were proposed to address this limitation [2].

Short texts and embedding-based models. Sparse word co-occurrence in short texts has motivated methods designed specifically for short documents [2]. More recently, embedding-based models have used contextual representations to capture semantic relationships beyond exact word overlap. BERTopic combines document embeddings, dimensionality reduction, density-based clustering, and c-TF-IDF topic representations [4], with sentence-transformer models supporting multilingual semantic representations [7, 11].

Topic-model comparison and evaluation. Automatic coherence measures such as C_v are designed to approximate human judgments of topic quality [12], while topic diversity measures the proportion of distinct words across topic representations. These metrics capture different aspects of model quality and should be considered together with topic coverage and substantive interpretability.

3. Materials And Methods

This study employed a comparative approach to analyze research topic trends using LDA and BERTopic. The analysis consisted of data collection, preprocessing, topic modeling, model evaluation, and trend analysis. Both methods were applied to the same journal title corpus to compare their topic structures, document coverage, and ability to identify temporal research patterns.

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

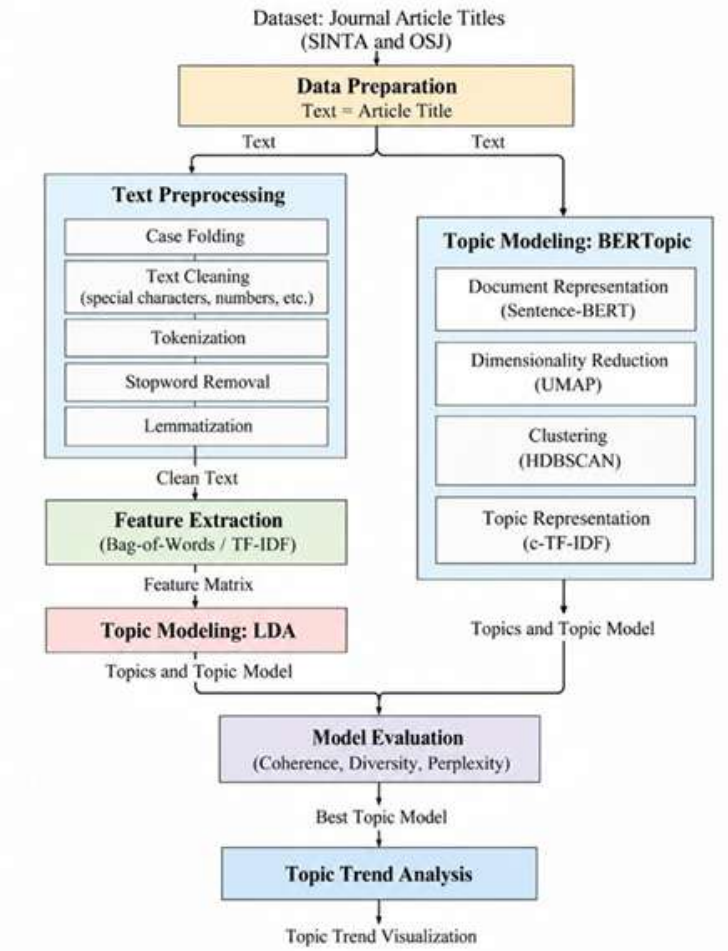


Figure 1. Data analysis workflow, from journal titles through LDA and BERTopic modeling to research trend analysis.

3.1. Dataset

Dataset was obtained from the OSJ website (OSJ.my.id), which provides journal article metadata, including journal names, SINTA rankings (S1 between S6), issue information, publication years, article titles, and article URLs.

The raw dataset contained 39,610 records. During data cleaning, 39 records that did not represent research articles, such as entries containing only journal names, and 39 duplicate records with identical titles and URLs were removed. This process resulted in a final dataset of 39,532 article titles from 283 journals. In total, 78 records, equivalent to 0.20% of the initial dataset, were removed. Four titles associated with different URLs were retained as separate articles.

For modeling, only the article title text was used as input. Publication year, journal name, and SINTA ranking were retained for temporal analysis and subgroup comparisons. The dataset covered the publication period from 2015 to 2025. However, 2025 represented an incomplete publication year, with 902 article titles from 43 journals. After cleaning, the article titles had an average length of 12.3 words, a median of 12 words, and a range of 3–36 words.

3.2. Language Handling

The corpus comprised article titles in two original languages: Indonesian, accounting for 21,266 titles (53.8%), and English, accounting for 18,266 titles (46.2%). In the raw dataset, the translated titles differed from the original titles in 15,973 of the 18,311 records originally written in English, representing 87.2% of those records. Rather than modeling the two languages separately, we adopted a monolingual approach in which article titles were modeled in Indonesian, while the original titles were retained in a separate column.

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Because translation can alter lexical choices and potentially affect topic representations, the original language was considered an important factor in evaluating the robustness of the findings. Furthermore, original language was strongly associated with SINTA ranking, as discussed in Section 4.8. Therefore, the robustness of the main findings was examined separately across strata defined by the original language.

3.3. Text Preprocessing

Punctuation was removed before tokenization. Stopwords were removed using a combined list of 217 unique words, comprising 19 Indonesian words and 198 English words. Twelve technical terms, including machine learning, deep learning, and decision support system, were protected from removal by preserving them as n grams. Stemming was not applied because article titles are relatively short, morphological variations may carry meaningful information, and overly aggressive normalization could remove distinctions relevant to topic identification.

After stopword removal, the corpus contained 424,890 tokens, averaging 10.8 tokens per title, and 30,583 unique unigrams. For both modeling approaches, the vocabulary consisted of unigrams, bigrams, and trigrams generated using scikit-learn's CountVectorizer, with a maximum document frequency of 0.85, a minimum document frequency of 5, and a maximum of 5,000 features [13]. This configuration ensured that both models used the same set of 5,000 terms to describe the identified topics.

3.4. Latent Dirichlet Allocation

LDA was implemented using scikit-learn with batch variational Bayes, 20 iterations, and the default prior of (1/K), applied to a bag-of-words matrix of size 39,532 × 5,000. The number of topics was evaluated across ($K \in$ 5,7,10,12,15,20,25,30),using three random seeds (42, 123, and 2024). This range extended the initial plan of 5–20 topics because preliminary experiments showed that coherence continued to increase at (K = 20).

The optimal number of topics was selected using a criterion defined before examining the results: the (K) value with the highest sum of the z-scores of mean coherence and mean topic diversity across the tested values of (K). Perplexity was reported as an additional evaluation metric but was not used for model selection. Each article title was assigned to the topic with the highest topic probability. Topic names were assigned by the authors after examining the ten most representative terms and representative article titles for each topic. These names serve as interpretive labels rather than outputs generated directly by the model.

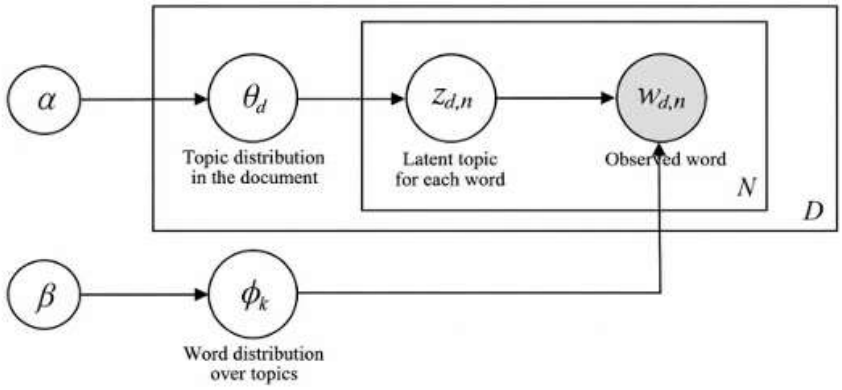


Figure 2. LDA graphical model (plate notation). Documents (D), words (N), and topics (K) are shown; only w_d,n is observed.

3.5. BERTopic

Article titles were converted into numerical representations using the multilingual sentence encoder paraphrase-multilingual-MiniLM-L12-v2 [7, 11]. The model was applied to cleaned titles with stopwords retained because sentence encoders are designed to capture the meaning of natural language. This encoder was selected because Indonesian article titles often contain English technical terms and the model can run on a free GPU environment. Other encoders were not tested, as discussed in Section 6.

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

The resulting embeddings were reduced using UMAP with 15 neighbors, five components, a minimum distance of 0, and the cosine metric [8]. HDBSCAN was then used to group similar embeddings, with a minimum cluster size of 50, minimum samples set to the same value, the Euclidean metric, and the excess of-mass cluster selection method [9]. Each topic was described using class-based TF-IDF (c-TF-IDF) and the same vocabulary of 5,000 terms used for LDA. The ten most representative terms were reported for each topic. The BERTopic language setting was changed to Indonesian because the default English setting removes non-alphanumeric characters, which can affect hyphenated terms.

UMAP was run with three random seeds: 42, 123, and 2024. Documents classified as outliers (topic −1) were retained and included in the results. To compare BERTopic with LDA using the same number of topics, the original BERTopic topics were reduced hierarchically to 25 topics. Sensitivity tests were also conducted using minimum cluster sizes of 30, 50, 100, and 200, along with the default minimum samples setting and a value of 10. These tests used seed 42.

All subsequent analyses, including topic alignment, trend analysis, and subgroup analysis, used the baseline BERTopic model with seed 42. This model assigned 24,551 article titles, or 62.1% of the corpus, to identified topics.

3.6. Evaluation

Topic coherence was measured using the Cv metric [12] implemented in Gensim [14]. The calculation used the ten most representative terms from each topic, along with the same n-gram texts and dictionary for both methods, containing 462,458 unique n-gram types. Topic diversity was calculated as the number of unique terms divided by the total number of top terms across all topics.

Topic size distributions were summarized using the median, range, Gini coefficient, and normalized entropy. Topic coverage was measured as the proportion of article titles assigned to a topic. Since BERTopic assigned topics to only part of the corpus, we also calculated LDA topic size statistics using only titles with a highest topic probability of at least 0.475. This restriction provided a coverage level comparable to that of BERTopic.

Interpretability was not evaluated by human raters. Although a blind evaluation form had been prepared, the assessment had not yet been conducted. Therefore, interpretability is reported only as a qualitative assessment by the authors.

3.7. Topic alignment

Because the two methods produced different sets of topics, topics were not matched based on their indices. Instead, three similarity measures were calculated for each LDA and BERTopic pair: (i) cosine similarity between the embedding centroids of the titles assigned to each topic; (ii) Jaccard overlap between the sets of titles assigned to each topic, considering only titles assigned by BERTopic; and (iii) Jaccard overlap between the sets of terms in the ten most representative phrases for each topic.

A pair was considered a strong match when the two topics were each other's most similar counterpart based on cosine similarity, and either the document overlap was at least 0.15 or the term overlap was at least 0.20. These thresholds were defined before examining the results and were used as practical conventions. Since all titles concerned computing, cosine similarity between topic centroids was high for almost every pair, ranging from 0.84 to 0.99. Therefore, cosine similarity was used for ranking candidate matches rather than as the sole evidence of topic correspondence.

An LDA topic without a strong match was classified as split if at least 30% of its titles assigned to BERTopic fell into a single BERTopic topic. Otherwise, it was classified as having no clear counterpart. Agreement between the two topic partitions was measured using normalized mutual information and the adjusted Rand index [15, 16], based on the 24,551 titles assigned by both methods.

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

3.8. Analisis Tren

For each topic, we calculated its annual proportion of titles. For LDA, the proportion was calculated using all titles published in a given year. For BERTopic, only titles assigned to topics were included in the denominator. This allowed us to compare topic proportions across the two methods. We estimated trends for 2015–2024 and excluded 2025 because the data for that year were incomplete and covered fewer journals. One journal accounted for 24.5% of all titles in 2025.

We examined trends using three methods: ordinary least squares (OLS) regression, Kendall's (\tau) with the Theil-Sen slope estimator [17, 18], and weighted least squares (WLS). For WLS, annual title counts were used as weights because the number of titles varied considerably, from 555 in 2015 to 8,536 in 2024.

Topics were classified into five categories using predefined criteria. A topic was classified as Emerging if its slope exceeded +0.3 percentage points per year with ($p < 0.10$), or Declining if its slope was below −0.3 percentage points per year with ($p < 0.10$). A topic was classified as New/Recent if its proportion was close to zero during the first three years before appearing and increasing. Cyclical topics had a coefficient of variation above 0.5 without a statistically significant trend. All remaining topics were classified as Persistent. We also reported Benjamini-Hochberg-adjusted (q)-values [19] for the OLS (p)-values. These thresholds were defined for this study and should not be interpreted as universal standards.

To determine whether the observed trends were affected by changes in the journal collection, we conducted six additional checks. First, we analyzed a panel of 96 journals that published articles in at least 8 of the 10 years. This panel contained 18,288 titles, representing 46.3% of the corpus. Second, we repeated the analysis after excluding IJASEIT, the largest journal, which contributed 2,875 titles. Its share of the annual corpus fell from 22.7% in 2015 to 3.0% in 2024, although its annual publication count increased from 126 to 256 titles. Third, we restricted LDA to titles assigned with the highest confidence. Fourth, we recalculated BERTopic proportions using all titles, including outliers, as the denominator. Fifth, we standardized the proportions to a fixed SINTA composition consisting of S1–S2, S3, and S4–S6 journals to account for changes in accreditation level composition [20]. Finally, we examined Indonesian-language and English-language titles separately and calculated language-standardized proportions.

We classified a topic as composition-sensitive if its standardized slope had the opposite sign to its raw slope and at least one had an absolute magnitude of 0.15 percentage points per year or greater. A topic was also flagged if more than half of its slopes with an absolute magnitude of at least 0.3 percentage points per year fell below that threshold after standardization. A topic was classified as language-specific if the slopes for Indonesian and English titles had opposite signs, with at least one reaching an absolute magnitude of 0.3 percentage points per year, or if the smaller absolute slope was less than 30% of the larger one.

We used two evidence tiers to summarize the robustness of the trends. Tier 1 included Emerging and Declining trends for which all three statistical tests agreed on the same category. The category also had to remain unchanged after IJASEIT was excluded. In addition, the SINTA-standardized slope had to retain the same sign with ($p < 0.10$), and the slopes across all three SINTA groups had to retain the same sign without being flagged as composition-sensitive. The trend also had to be consistent across languages rather than language-specific.

Tier 2 included trends for which all three tests agreed on the direction and statistical significance (($p < 0.10$)), but assigned different categories because a slope fell just below the predefined magnitude threshold. All other robustness criteria had to remain satisfied.

3.9. Subgroup Analysis

SINTA rankings were grouped into three categories for S1–S2 (20 journals), S3 (43 journals), and S4–S6 (220 journals). The S6 category alone accounted for 594 titles from nine journals. These rankings reflect the journals' current accreditation status, not their status at the time the articles were published.

Because titles from the same journal cannot be assumed to be independent, we did not rely on chi-square (p) values. Instead, associations between topics and SINTA groups or language were summarized using Cramér's V [21].

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Differences in topic proportions between groups were reported with 95% percentile confidence intervals based on 2,000 bootstrap resamples of journals within each group.

We also examined how titles were distributed across journals within each topic. Journal concentration was measured using the share of the largest contributing journal and the effective number of journals, calculated as the inverse of the sum of squared journal shares. Finally, journal distinctiveness was assessed using Jensen-Shannon distance [22] between each journal's topic profile and the overall corpus profile.

3.10. Implementation and Reproducibility

All analyses were conducted in Python on Google Colab using scikit-learn, Gensim, BERTopic, umap-learn, HDBSCAN, SciPy, and statsmodels [13, 14, 23, 24]. [Specify the exact package versions used in the runtime.] Random seeds, model parameters, and all result tables were saved alongside the analysis notebook, and every reported value can be traced to the corresponding output files (Supplementary Table S1). The code and derived data are available in the GitHub repository at https://github.com/Zamy17/lda-bertopic-cs-trends.

Table 1. Configurations of the two topic models.

Component	Configuration
Corpus	39.532
Vocabulary	Unigram to trigrams; maximum document frequency = 0.85, minimum document frequency = 5; 5,000 features; shared by both methods
LDA	scikit-learn LatentDirichletAllocation; batch variational Bayes; 20 iterations; prior = 1/K; K ∈ {5, 7, 10, 12, 15, 20, 25, 30}; random seeds = 42, 123, 2024; model selection based on the combined z-score of coherence and diversity
Embedding BERTopic	paraphrase-multilingual-MiniLM-L12-v2 applied to cleaned titles
UMAP	n_neighbors = 15; n_components = 5; min_dist = 0; cosine metric; random seeds = 42, 123, 2024
HDBSCAN	Minimun cluster size = 50 (baseline; 30, 100, and 200 tested); min_samples = minimum cluster size (10 also tested); Euclidean metric; excess-of-mass cluster selection
Topic Representation	Class-based TF-IDF TF-IDF (c-TF-IDF) using the shared vocabulary; 10 representative terms per topic; language setting = "indonesian"
Topic Reduction	Hierarchical reduction to 25 topics, with outliers retained
Evaluation	C_v coherence (Gensim) and topic diversity based on the top 10 terms; identical n-gram texts and dictionary for both methods

Further analyses use the baseline model with random seed 42 (minimum cluster size 50).

4. Results

4.1. Corpus Description

The corpus grew from 555 titles in 2015 to 8,536 in 2024, a 15.4-fold increase. The number of contributing journals also increased, reaching 265 in 2022–2023 (Table 2, Figure 1). IJASEIT, the largest journal in the corpus, accounted for 22.7% of titles in 2015 but only 3.0% in 2024.

The proportion of titles originally written in English ranged from 39.1% to 56.4% between 2015 and 2024. The 2025 data comprised only 902 titles from 43 journals. One journal contributed 24.5% of these titles. Although 2025 is included in the descriptive results, it was excluded from slope estimation because the year's data were incomplete.

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Table 2. Corpus composition by year.

Year	Title	Journals	Largest Journal Share (%)	Originally English Titles (%)	Outlier BERTopic (%)
2015	555	46	22.7	56.4	41.4
2016	929	80	19.9	46.1	39.8
2017	1,548	108	20.5	47.7	40.2
2018	1,966	135	18.8	43.0	39.2
2019	2,585	171	11.8	40.0	39.1
2020	3,883	223	9.2	43.5	38.8
2021	4,873	252	6.8	39.1	36.9
2022	5,894	265	5.6	43.3	36.7
2023	7,861	265	8.4	49.5	36.9
2024	8,536	252	7.8	50.7	38.0
2025	902	43	24.5	60.1	41.5

Originally English titles: titles originally written in English and translated into Indonesian for modeling. BERTopic outliers: proportion of titles labeled −1 in the seed-42 model.

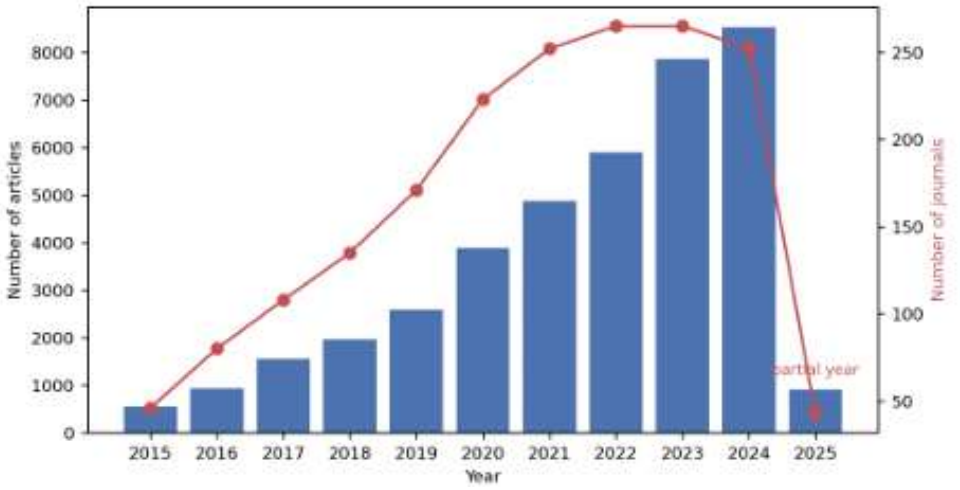


Figure 3. Number of titles (bars) and contributing journals (line) by year. 2025 is an incomplete year.

4.2. LDA Model Selection

Coherence increased from 0.532 (SD = 0.011) at (K = 5) to approximately 0.58 at (K = 20) and (K = 25), with scores of 0.581 (SD = 0.017) and 0.580 (SD = 0.039), respectively. It then decreased to 0.563 (SD = 0.012) at (K = 30) (Table 3, Figure 2). Diversity fluctuated between 0.66 and 0.71 without a consistent upward or downward trend, while training-set perplexity decreased as (K) increased, from 1,174 at (K = 5) to 859 at (K = 25) and 858 at (K = 30).

The predefined combined z-score was highest at (K = 25) (+1.43), followed by (K = 20) (+0.94). However, the difference between these two values was small relative to the variation across random seeds. At (K = 25), coherence ranged from 0.537 to 0.611 across seeds. We therefore interpreted (K = 20)–25 as a plateau rather than a clearly defined optimum and retained (K = 25), as planned.

Removing 17 generic academic terms, such as menggunakan, metode, and berbasis, increased diversity at (K = 25) by 0.113 but reduced coherence by 0.020. Since (C_v) was calculated separately for each corpus configuration, these scores are not fully comparable. We therefore retained the baseline configuration for the main analysis.

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Table 3. LDA model selection (mean and standard deviation across three seeds).

K	Coherence (C_v)	Topic diversity	Perplexity	Combined z
5	0.532 (SD 0.011)	0.713 (SD 0.023)	1174	-0.01
7	0.544 (SD 0.011)	0.714 (SD 0.043)	1107	+0.68
10	0.534 (SD 0.015)	0.687 (SD 0.032)	1025	-1.39
12	0.553 (SD 0.030)	0.700 (SD 0.008)	964	+0.35
15	0.554 (SD 0.031)	0.660 (SD 0.018)	939	-1.83
20	0.581 (SD 0.017)	0.683 (SD 0.006)	888	+0.94
25	0.580 (SD 0.039)	0.693 (SD 0.008)	859	+1.43
30	0.563 (SD 0.012)	0.681 (SD 0.015)	858	-0.18

Combined z-score: the sum of the z-scores of mean coherence and mean diversity across all \(K\) values (selection criterion). Perplexity was calculated on the training data.

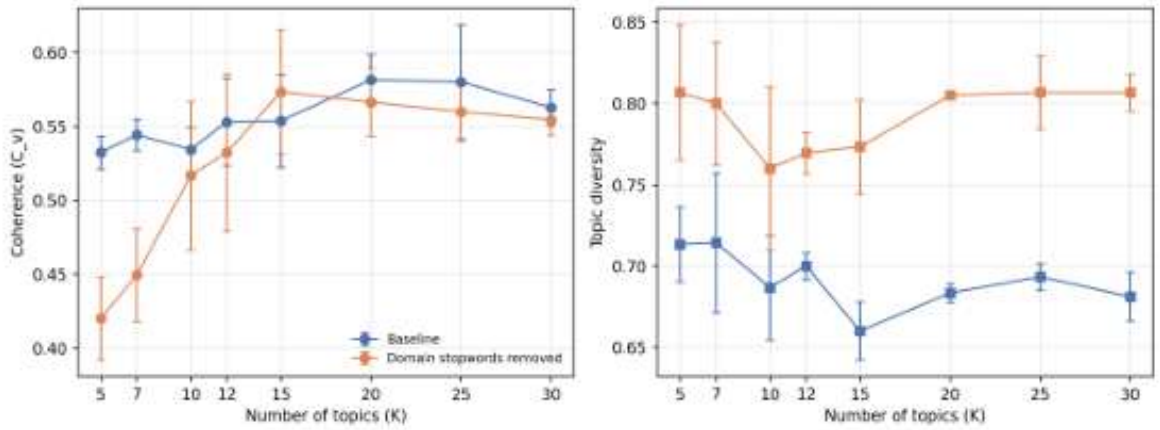


Figure 4. LDA coherence (C_v) and topic diversity by number of topics (mean and standard deviation across three seeds), for baseline preprocessing and preprocessing with 17 generic terms removed.

4.3. BERTopic

The baseline configuration produced 121 topics, with 120–122 topics across random seeds, and classified 38.45% of titles as outliers (SD = 0.49%) (Table 4). Coherence was 0.721 (SD = 0.005), while diversity was 0.814 (SD = 0.005).

The outlier rate was relatively stable across years, ranging from 36.7% to 41.5%, but differed by original language. It was 35.6% for titles originally written in Indonesian and 40.6% for those originally written in English. Varying the clustering parameters produced 67–185 topics, outlier rates of 33.6%–41.1%, and coherence scores of 0.706–0.729 for seed 42. No non-degenerate configuration reduced the outlier rate below 33.6%.

A minimum cluster size of 200 resulted in a degenerate solution with only two topics, one of which contained 98.0% of all titles. This configuration was excluded from the conclusions. Hierarchical reduction to 25 topics did not change the number of outliers and yielded a coherence score of 0.671 (SD = 0.017) and a diversity score of 0.841 (SD = 0.016).

The two-dimensional projection (Figure 9) shows several dense clusters, including those associated with learning media and augmented reality, surrounded by a diffuse band of outliers.

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

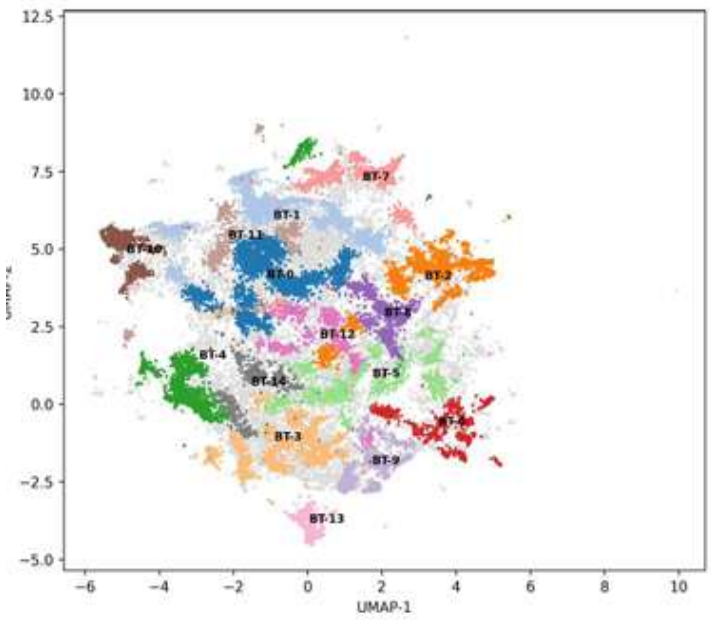


Figure 5. Two-dimensional UMAP projection of title embeddings, colored by the 15 largest topics in the 25 topic BERTopic solution (gray: outliers). BT-n labels correspond to B-n in Table 6.

4.4. Topic Inventory (RQ1)

Both methods identified similar broad themes, including information systems and applications, decision support systems, learning media, the Internet of Things, networking and security, and machine learning applications (Tables 5 and 6).

The 25 LDA topics contained between 956 and 3,953 titles each, representing 2.4%–10.0% of the corpus. The largest topics were Web-Based Information Systems (10.0%), Machine Learning and Intelligent Detection (8.7%), and Decision Support Systems (5.4%). In comparison, the 25 BERTopic topics ranged from 73 to 3,166 titles, representing 0.3%–12.9% of the titles assigned to topics.

BERTopic also identified several narrower themes that LDA did not separate at (K = 25). These included antennas (124 titles), chatbots (123), blockchain (122), cryptography (470), COBIT governance (291), the waterfall method (263), two COVID-19 topics (247 and 194), and immigration services (73).

Some topics were at the boundaries of computer science. These included Banking, Finance, and Accounting (B7, 1,237 titles) and Plant Image Classification and Agricultural/Food Studies (B9, 1,177 titles). Several LDA topics had top terms dominated by generic words, such as menggunakan (using) and metode (method). Other LDA topics combined multiple themes. For example, the Internet of Things topic (L13) also contained COVID 19-related terms, while L8 combined nearest-neighbor terminology with terms related to electrical energy.

Topic names were assigned as interpretive labels after examining representative titles

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Table 4. LDA topics (K = 25), sorted by size.

ID	Topic Label	Title Of Number (%)	Top terms
L15	Web-Based Information Systems	3,953 (10.0)	Sistem, informasi, sistem informasi, berbasis, web, berbasis web
L10	Machine Learning and Intelligent Detection	3,425 (8.7)	Jaringan, menggunakan, pembelajaran, analisis, pendekatan, berbasis
L17	Sdecision support Sytems	2,120 (5.4)	Metode, keputusan, sistem, pendukung, pendukung keputusan, sistem pendukung
L20	Prediktive Clasification and Sentiment Analysts	1,968 (5.0)	Algoritma, perbandingan, menggunakan, bayes, menggunakan algoritma, analisis
L12	UI UX and Mobile Application	1,836 (4.6)	Aplikasi, menggunakan, metode, perancangan, mobile, menggunakan metode
L3	Microcontrollers and Automational Strategy	1,705 (4.3)	Sistem, menggunakan berbasis, alat, implementasi, perancangan
L11	Employee perpormance and Organization Strategy	1,647 (4.2)	Teknologi, kinerja, kerja, tinggi, strategi, digital
L22	Deep Learning dan Convolutional Neural Network	1,628 (4.1)	Menggunakan, klasifikasi, network, neural, neural network, vector
L8	Elektrical Energy and power Sytems	1,552 (3.9)	Nearest, listrik, neighbor, nearest neighbor, analisis, menggunakan
L7	Data Mining and Pattern Anaylst	1,550 (3.9)	Data, menggunakan, means,algoritma, clustering, mining
L19	Consumer Behavior and Servis Quality	1,503 (3.8)	Studi, kasus,studi kasus, faktor, analisis kualitas
L16	Android and Mobile Applications	1,409 (3.6)	Aplikasi, berbasis, android, berbasis android, implementasi, pengembangan
L9	Augmented Reality and Educational Games	1,365 (3.5)	reality, game, augmented, augmented reality, berbasis, perangkat
L21	Fuzzy Logic and Image Processing	1,363 (3.4)	Fuzzy,metode, menggunakan, citra, produksi, menggunakan metode
L5	Planning and information Systems	1,288 (3.3)	Bangun, rancang, rancang bangun, sistem, informasi, bangun sistem
L23	Sentiment Analyst and Social Media	1,266 (3.2)	Analisis, menggunakan, sosial, sentimen, indonesia, media
L1	Learning Media and Eduacation	1,218 (3.1)	Pembelajaran, siswa, sekolah, media, negeri, berbasis
L13	Internet of Things and Monitoring Sytems	1,176 (3.0)	Internet, covid 19, iot, berbasis
L18	Finance and Corporate Performance	1,153 (2.9)	Perusahaan, pengaruh, indonesia, analisis, tahun,nilai
L6	Artifical Neural Networks and Optimizations	1,140 (2.9)	Algoritma, menggunakan, optimasi, fitur, jaringan, menggunakan algoritma
L0	User Statification and Usability	1,121 (2.8)	Menggunakan, metode, pengguna, analisis, menggunakan metode, kepuasan
L2	Information Technology and Govemance And Security	1,105 (2.8)	Menggunakan, informasi, teknologi, keamanan, teknologi informasi; framework
L14	Experts Systems ans Diagnosis	1,071 (2.7)	sistem; metode, penyakit, pakar, sistem pakar, menggunakan
L24	Learning and Education	1,014 (2.6)	Mahasiswa, universitas, belajar, program, teknik, hasil
L4	**SAW-** Based Decision Support Systems	956 (2.4)	Penerapan, penerapan metode, metode, algoritma, simple, penerapan algoritma

Table 5. BERTopic topics after reduction to 25, sorted by size (seed 42; percentages are based on 24,551 assigned titles).

ID	Topic Label	Number of Title (%)	Top Terms
B0	Web-Based Information Systems and Applications	3,166 (12.9)	sistem informasi, informasi, sistem, berbasis, web, berbasis web
B1	Learning Media And E-Learning	2,427 (9.9)	Pembelajaran, siswa, berbasis, media, belajar, media pembelajaran
B2	Sdecision Support Systems and Multi-Criteria Methods	2,109 (8.6)	Keputusan, pendukung keputusan, pendukung, sistem pendukung, sistem pendukung keputusan, metode
B3	IoT, Microcontrollers, and Monitoring Systems	1,946 (7.9)	Air, robot, ikan, sistem, berbasis, listrik
B4	Computer Networks and Network Security	1,681 (6.8)	Jaringan, keamanan, cloud, menggunakan, nirkabel, analisis
B5	Swarm Intelligence, Optimization, and Applied Machine Learning	1,656 (6.7)	Algoritma, clustering, means, klasifikasi, menggunakan, means clustering
B6	Medical Image Analysis and Disease Diagnosis	1,261 (5.1)	Penyakit, wajah, pakar, sistem pakar, klasifikasi, diabetes
B7	Banking, Finance, and Accounting	1,237 (5.0)	Pengaruh, perusahaan, keuangan, akuntansi, pajak, kerja
B8	Sentiment Analysis and Social Media	1,210 (4.9)	Sentimen, analisis sentimen, twitter, sosial, media sosial, analisis
B9	Plant Image Classification and Agricultural/Food Studies	1,177 (4.8)	Tanaman. Penyakit, daun, buah, padi, penyakit tanaman
B10	Augmented Reality and Educational Games	1,085 (4.4)	Reality, augmented, augmented reality, game, edukasi, media
B11	UI/UX, Design Thinking, and Extreme Programming	1,010 (4.1)	Design, perpustakaan, design thinking, thinking, ux, user
B12	Business Prediction and Forecasting (Regression, Time Series, Apriori)	815 (3.3)	Algoritma,penjualan,algoritma apriori, kopi, pola
B13	Fuzzy Logic and Inference Systems	661 (2.7)	Fuzzy, metode fuzzy, logika fuzzy, logika, menggunakan metode fuzzy, tsukamoto
B14	Vehicles, Traffic, and Transportation IoT	626 (2.5)	Sampah, kendaraan, parkir, lalu lintas, lalu, lintas
B15	Information Systems and Healthcare Service Queues	577 (2.4)	Obat, sakit, rumah sakit, kesehatan, rumah, apotek
B16	Cryptography and Steganography	470 (1.9)	Enkripsi, kriptografi, cipher, keamanan, algoritma, pengamanan
B17	IT Governance and Auditing (COBIT)	291 (1.2)	Cobit, framework cobit, kelola, tata kelola, menggunakan framework cobit, tata
B18	The Waterfall Method in Systems Development	263 (1.1)	Waterfall, metode waterfall, menggunakan metode waterfall, waterfall sistem,model waterfall, sistem informasi
B19	COVID-19 Pandemic: IT Adoption and Strategies	247 (1.0)	Pandemi, pandemi covid,pandemi
B20	COVID-19: Forecasting, Modeling, and Sentiment Analysis	194 (0.8)	19, covid 19, covid, vaksin, covid 19 menggunakan, 19 menggunakan
B21	Antennas and Wireless Communications	124 (0.5)	Antena, mikrostrip, antena mikrostrip, patch, array, ghz
B22	Chatbots and Natural Language Processing	123 (0.5)	Chatbot, layanan, layanan informasi, language, rasa, chat
B23	Blockchain Technology	122 (0.5)	Blockchain, teknologi blockchain, berbasis blockchain, teknologi, keamanan; kontrak
B24	Immigration Services (Single Journal)	73 (0.3)	Imigrasi, kantor imigrasi, kantor imigrasi kelas, imigrasi kelas, tpi, kantor

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

4.5. Comparative Evaluation (RQ3)

With the number of topics standardized at 25, BERTopic achieved higher coherence than LDA (0.671, SD = 0.017, versus 0.580, SD = 0.039) and higher diversity (0.841, SD = 0.016, versus 0.693, SD = 0.008). The mean ± SD ranges for both metrics did not overlap across the three random seeds (Table 4, Figure 3).

However, this advantage came with lower coverage. BERTopic assigned topics to 61.5% of the titles and left 37.9% unassigned, whereas LDA assigned every title to a topic. Topic sizes were also more evenly distributed in LDA, with a Gini coefficient of 0.20, normalized entropy of 0.98, and topic sizes ranging from 956 to 3,953 titles. BERTopic had a Gini coefficient of 0.45, normalized entropy of 0.90, and topic sizes ranging from 73 to 3,166 titles (Figure 4).

Restricting LDA to its highest-confidence assignments, reducing its coverage to 62.5% to match BERTopic, increased the Gini coefficient only slightly, to 0.25. Interpretability was not evaluated by human raters.

Table 6. Comparison of LDA and BERTopic.

Method	Topic	Coherence (C_v)	Diversity	Coverage (%)	Median Topic Size	Topic Size Range	Gini Coefficient	Norm. entropy
LDA (K=25)	25	0.580 (SD 0.039)	0.693 (SD 0.008)	100.0	1365	956–3953	0.20	0.98
BERTopic, reduced to 25	25	0.671 (SD 0.017)	0.841 (SD 0.016)	61.5	815	73–3166	0.45	0.90
BERTopic, natural	121	0.721 (SD 0.005)	0.814 (SD 0.005)	61.5	123	50–1843	0.46	0.92
LDA (K=25), coverage-matched	25	n/a	n/a	62.1	778	505–2832	0.25	0.96

Coherence, diversity, and coverage: means across three random seeds (with standard deviations). Topic-size statistics: calculated from the seed-42 model using assigned titles; coverage for this model was 62.1%. LDA perplexity ($(K = 25)$): 859. Coverage-matched LDA: titles with a topic-assignment probability (≥ 0.475); coherence and diversity were not recalculated.

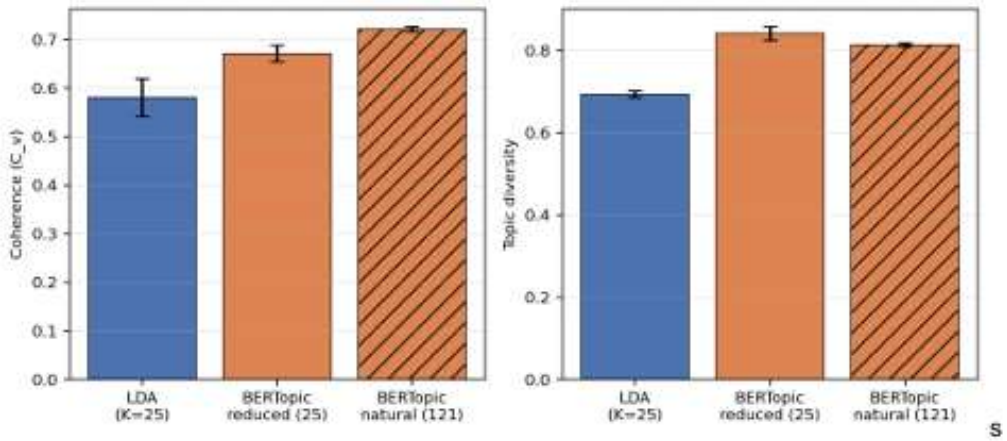


Figure 6. Coherence and diversity for LDA (K = 25), BERTopic reduced to 25 topics, and the natural BERTopic solution (mean of 121 topics). Bars represent means; whiskers represent standard deviations across three seeds. BERTopic covers 62.1% of titles; LDA covers 100%.

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

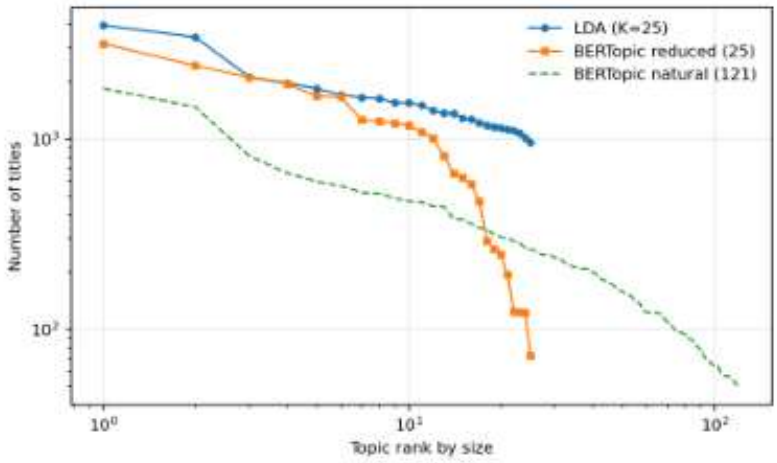


Figure 7. Rank-size distribution of topics (log-log scale).

4.6. Topic Alignment (RQ2)

Eleven of the 25 LDA topics had strong one-to-one matches with BERTopic topics (Table 7). The clearest matches were Augmented Reality and Educational Games (cosine similarity = 0.985, document Jaccard = 0.534), Decision Support Systems (0.435), Sentiment Analysis and Social Media (0.324), IT Governance (0.312), Web-Based Information Systems (0.309), and UI/UX (0.293).

Not all strong matches represented closely aligned themes. For example, Electrical Energy and Power Systems was matched with IoT, Microcontrollers, and Monitoring Systems, despite a document Jaccard score of only 0.149. Such matches should therefore be interpreted with caution.

Of the 14 LDA topics without a strong match, nine were classified as split topics, meaning that at least 30% of their titles assigned by BERTopic fell into a single BERTopic topic. The remaining five topics, L3, L13, L16, L20, and L22, had no clear counterparts. Document overlap also revealed an uneven correspondence between the two methods. The 25 LDA topics mapped predominantly to only 13 BERTopic topics, while 12 BERTopic topics, covering 4,295 titles or 17.5% of all titles assigned by BERTopic, were not the dominant counterpart of any LDA topic (Figure 5).

The two LDA topics with the steepest increases, Deep Learning and Convolutional Neural Networks (L22) and Predictive Classification and Sentiment Analysis (L20), had no clear BERTopic counterparts. Their titles were distributed across several application-oriented BERTopic topics. For L22, the largest shares were Medical Image Analysis (25.8%), Plant Image Classification (20.6%), Swarm Intelligence and Applied Machine Learning (15.1%), and Sentiment Analysis (12.5%). For L20, the titles were distributed across Sentiment Analysis (29.1%), Swarm Intelligence and Applied Machine Learning (14.5%), Medical Image Analysis (13.3%), and Decision Support Systems (13.2%).

Overall, the two methods showed moderate agreement across the 24,551 titles assigned by both methods, with a normalized mutual information (NMI) of 0.314 and an adjusted Rand index (ARI) of 0.181. When the original 121-topic BERTopic solution was used, the NMI increased to 0.375.

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Table 7. Strong LDA–BERTopic topic pairs (mutual best cosine and document Jaccard ≥ 0.15 or word Jaccard ≥ 0.20), sorted by document overlap.

LDA Topic	BERTopic	Cosine	Doc-Jaccard	Word-Jaccard
L9:Augmented Reality and Educational Games	B10: Augmented Reality and Educational Games	0.985	0.534	0.455
L17: Decision Support Systems	B2: Decision Support Systems and Multi-Criteria Methods	0.974	0.435	0.556
L23: Sentiment Analysis and Social Media	B8: Sentiment Analysis and Social Media	0.953	0.324	0.500
L2: Tata Kelola dan Keamanan Teknologi Informasi	B17: IT Governance and Auditing (COBIT)	0.902	0.312	0.364
L15: Web-Based Information Systems	B0: Web-Based Information Systems and Applications	0.961	0.309	0.750
L12: UI/UX and Mobile Applications	B11: UI/UX, Design Thinking, and Extreme Programming	0.958	0.293	0.333
L18: Finance and Corporate Performance	B7: Banking, Finance, and Accounting	0.922	0.282	0.250
L10: Machine Learning and Intelligent Detection	B4: Computer Networks and Network Security	0.872	0.231	0.176
L1: Learning Media and Education	B1: Learning Media and E-Learning	0.968	0.220	0.357
L14: Expert Systems and Diagnosis	B6: Medical Image Analysis and Disease Diagnosis	0.883	0.155	0.385
L8: Electrical Energy and Power Systems	B3: IoT, Microcontrollers, and Monitoring Systems	0.888	0.149	0.267

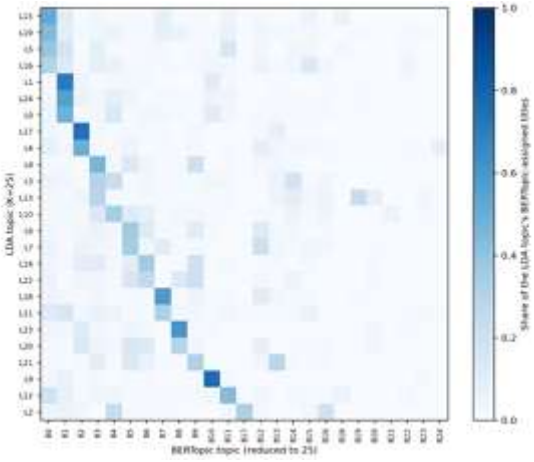


Figure 8. Document-level overlap between LDA and BERTopic topics: proportion of each LDA topic's assigned titles that fall into each BERTopic topic (L: LDA, B: BERTopic; identifiers correspond to Tables 5 and 6).

4.7. Temporal Trends (RQ4)

Under the primary OLS classification, LDA identified five Emerging topics, four Declining topics, and 16 Persistent topics. BERTopic identified three Emerging topics, two Declining topics, five New/Recent topics, and 15 Persistent topics (Figures 6 and 7). All three statistical tests agreed on the category assigned to 21 of the 25 LDA topics and 23 of the 25 BERTopic topics. Table 8 summarizes the Emerging, Declining, and New/Recent topics, together with their levels of support after the robustness checks described in Section 3.8.

Seven method-topic combinations met all Tier 1 criteria. For LDA, these were Predictive Classification and Sentiment Analysis (+0.55 percentage points per year), Deep Learning and Convolutional Neural Networks (+0.43), and the Internet of Things and Monitoring Systems (+0.43). For BERTopic, Sentiment Analysis and Social Media (+0.57), Medical Image Analysis and Disease Diagnosis (+0.37), and Swarm Intelligence, Optimization, and Applied Machine Learning (+0.35;(q = 0.053)) were classified as Emerging. Learning Media and E-Learning was classified as Declining (−0.48;(q = 0.096)).

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Two additional LDA trends met the directional and significance criteria across all three tests but did not retain the same category because their weighted slopes fell just below the threshold of 0.3 percentage points per year. These were the increase in Data Mining and Pattern Analysis and the decline in Augmented Reality and Educational Games. Both were classified as Tier 2.

The two COVID-19 topics increased in 2020 and peaked in 2021 and 2022 before declining to nearly zero by 2025 (Figure 7). Their New/Recent classification therefore reflects a rise associated with a specific event rather than sustained growth. By contrast, chatbot and blockchain topics, each accounting for approximately 0.5% of titles assigned by BERTopic, continued to increase through 2024. The immigration services topic was dominated by a single journal, which contributed 83.6% of its titles. It was therefore not interpreted as evidence of a broader research trend.

Among the 11 strong LDA–BERTopic topic matches, seven pairs had slopes with the same sign, while five received the same trend category. The median Spearman correlation between their annual proportions was 0.48, with values ranging from −0.64 to 0.96 (Figure 8).

Agreement was strongest for pairs with clearly aligned themes: Learning Media ((ρ = 0.96)), Sentiment Analysis ((ρ = 0.93)), UI/UX and Decision Support Systems ($both(\rho$ = 0.77)), and Augmented Reality ((ρ=0.75)). The weakest agreement occurred for a less closely aligned pair, Expert Systems and Diagnosis versus Medical Image Analysis ((ρ=−0.64)).

Table 8. Trend classification and level of support (slope: OLS, percentage points per year for 2015–2024; q: Benjamini–Hochberg).

Support	Method	Topic	Category	Slope	Q
Tier 1	LDA	Classification, Prediction, and Sentiment Analysis	Emerging	+0.55	0.000
Tier 1	LDA	Deep Learning dan Convolutional Neural Network	Emerging	+0.43	0.000
Tier 1	LDA	Internet of Things and Monitoring Systems	Emerging	+0.43	0.005
Tier 1	BERTopic	Sentiment Analysis and Social Media	Emerging	+0.57	0.002
Tier 1	BERTopic	Learning Media and E-Learning	Declining	−0.48	0.096
Tier 1	BERTopic	Medical Image Analysis and Disease Diagnosis	Emerging	+0.37	0.000
Tier 1	BERTopic	Swarm Intelligence, Optimization, and Applied Machine Learning	Emerging	+0.35	0.053
Tier 2	LDA	Augmented Reality and Educational Games	Declining	−0.37	0.005
Tier 2	LDA	Data Mining and Pattern Analysis	Emerging	+0.30	0.000
Not supported	LDA	Electrical Energy and Power Systems	Declining	−0.73	0.006
Not supported	LDA	Fuzzy Logic and Image Processing	Declining	−0.58	0.000
Not supported	LDA	Employee Performance and Organizational Strategy	Declining	−0.35	0.006
Not supported	LDA	UI/UX and Mobile Applications	Emerging	+0.31	0.005
Not supported	BERTopic	Plant Image Classification and Agricultural/Food Studies	Declining	−1.11	0.050
Event/recent	BERTopic	COVID-19 Pandemic: Information Technology Adoption and Strategy	New/Recent	+0.17	0.142
Event/recent	BERTopic	COVID-19: Forecasting, Modeling, and Sentiment Analysis	New/Recent	+0.14	0.087
Event/recent	BERTopic	Chatbots and Natural Language Processing	New/Recent	+0.09	0.002
Event/recent	BERTopic	Blockchain Technology	New/Recent	+0.07	0.007
Event/recent	BERTopic	Immigration Services (single journal)	New/Recent	+0.02	0.654

Tier 1: All criteria are satisfied. Tier 2: All three tests agree on the trend direction and statistical significance, but the assigned category differs only because of the magnitude threshold. Not supported: At least one robustness criterion is not met (the failed criterion is reported). Event/Recent: The topic is classified as New/Recent.

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

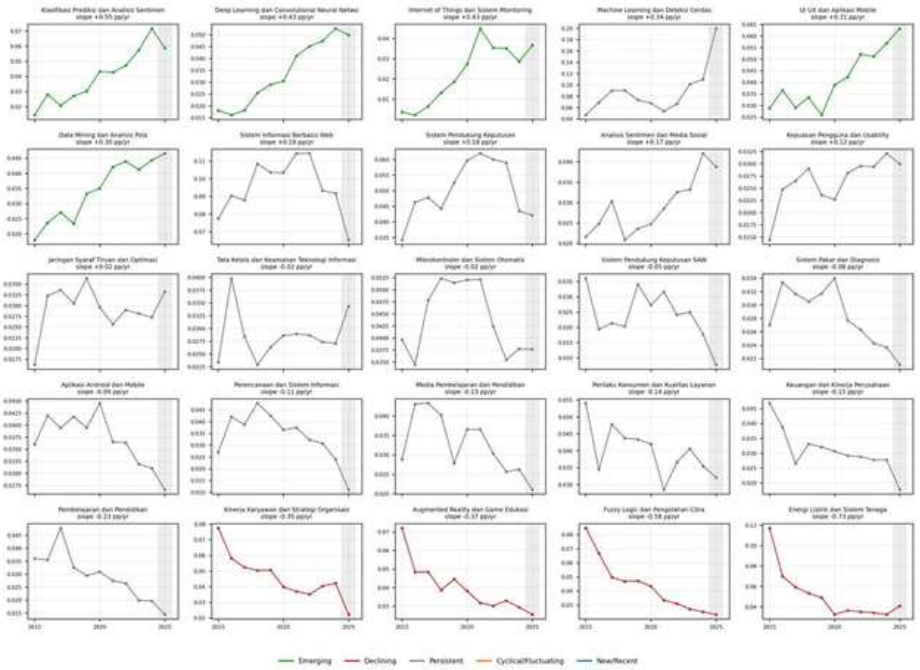


Figure 9. Annual proportion of titles for each LDA topic (color: trend category; gray band: incomplete 2025, excluded from slope estimation). Panels are ordered by slope; each panel has its own vertical scale.

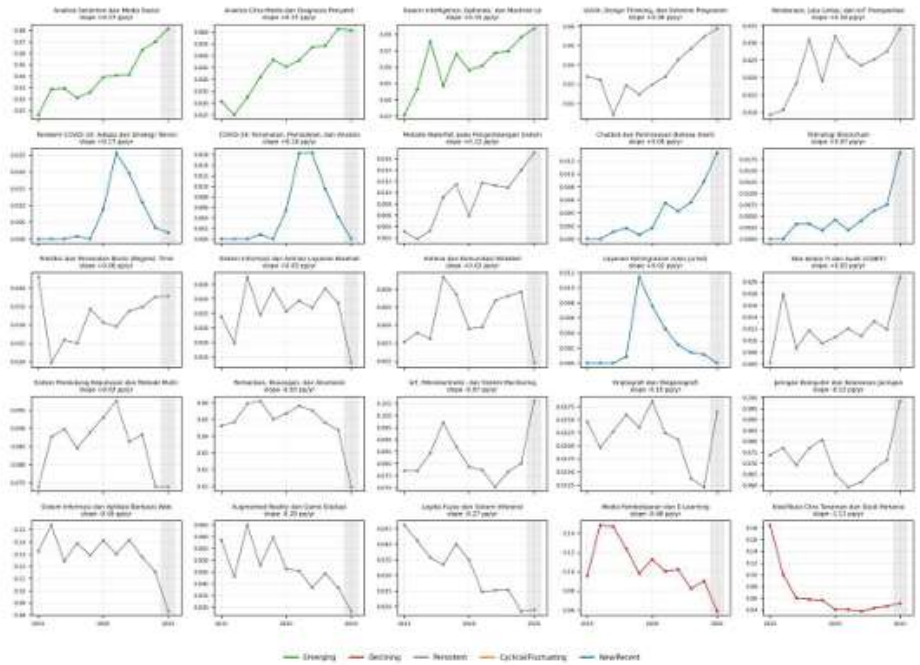


Figure 10. Annual proportion of assigned titles for each BERTopic topic (color: trend category; gray band: 2025). Panels are ordered by slope; each panel has its own vertical scale.

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)



Figure 11. Annual proportions for the 11 strongly aligned topic pairs (L: LDA, solid line; B: BERTopic, dashed line). ρ denotes Spearman correlation for 2015–2024.

4.8. Trend Robustness and Subgroup Differences

Label Trend classifications remained stable when the analysis was restricted to the panel of 96 journals. Across both methods, 96% of topics retained their original category and trend direction (Table 9). Excluding IJASEIT led to greater changes. The original category was retained for 76% of LDA topics and 84% of BERTopic topics, while the trend direction remained unchanged for 96% and 80%, respectively.

The steepest declines were sensitive to the inclusion of IJASEIT. For LDA, Electrical Energy and Power Systems had a slope of (−0.73) percentage points per year ((q = 0.006)). After IJASEIT was excluded, the slope was only (-0.04) percentage points per year ((p = 0.59)). IJASEIT accounted for 39.6% of the titles in this topic, with an effective journal count of 5.8. Similarly, BERTopic's Plant Image Classification and Agricultural/Food Studies topic had a slope of (-1.11) percentage points per year ((q = 0.050)), which changed to (+0.01) after IJASEIT was excluded. The journal contributed 31.1% of the titles in this topic.

After standardization to a fixed SINTA composition, the slopes for these two topics were (-0.32) ((p = 0.006))and (-0.54) ((p = 0.011)), respectively. Their trends also differed by original language. For Electrical Energy and Power Systems, the slope was (+0.01) for titles originally written in Indonesian and (1.38) for those originally written in English. For Plant Image Classification and Agricultural/Food Studies, the corresponding slopes were (0.00) and (-2.22). Both declines were concentrated in S1 to S2 journals. For Plant Image Classification and Agricultural/Food Studies, the slope was (-3.15) percentage points per year in S1 to S2 journals, compared with (+0.08) in S4 to S6 journals.

The distribution of SINTA groups also changed substantially over the study period. The share of titles from S4–S6 journals increased from 41.3% in 2015 to 68.0% in 2024, while the share from S1–S2 journals fell from 38.1% to 20.1%. These changes can affect the observed topic trends because the composition of the corpus varies over time.

This effect was evident for Web-Based Information Systems. The LDA topic had a raw slope of (+0.18) percentage points per year, but its standardized slope was (-0.08) ((p = 0.49)). The corresponding BERTopic topic had a raw slope of (-0.16), compared with a standardized slope of (-0.45) ((p = 0.002)). This topic was more prevalent in S4 to S6 journals than in S1 to S2 journals. Under BERTopic, it accounted for 16.3% of titles in S4 to S6 journals and 3.9% in S1 to S2 journals. Under LDA, the corresponding proportions were 13.6% and 1.7%. The changing share of S4 to S6 journals therefore masked the decline observed after standardizing the SINTA composition.

A different pattern appeared for LDA's Machine Learning and Intelligent Detection topic. Its raw slope was (+0.34) percentage points per year ((p = 0.13)), increasing to (+0.58) after standardization ((p = 0.003)). The increase was

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

concentrated in S1 to S2 journals, where the slope was (+2.74) percentage points per year, compared with (-0.03) in S4 to S6 journals. The trend also differed by original language: the slope was (+0.89) for titles originally written in English and (-0.21) for those originally written in Indonesian. This topic was therefore flagged as both composition-sensitive and language-specific.

Topic distributions differed systematically across SINTA groups (Table 10). Cramér's (V) between topic assignment and SINTA group was 0.34 for LDA and 0.24 for BERTopic. Differences in topic proportions between S1 aorund S2 and S4 S6 journals exceeded the 95% bootstrap confidence intervals for 16 of the 25 LDA topics and 18 of the 25 BERTopic topics.

S1–S2 journals had higher proportions of Machine Learning and Intelligent Detection (27.0% of titles, compared with 2.8% in S4 to S6 journals) and Electrical Energy (11.1% versus 1.9%) under LDA. Under BERTopic, IoT, Microcontrollers, and Monitoring Systems was also more prevalent in S1 between S2 journals (15.6% versus 6.0%). In contrast, S4 between S6 journals had a higher proportion of Web-Based Information Systems under LDA (13.6% versus 1.7%), as well as higher representation of Decision Support Systems, UI/UX, and Learning Media.

Table 9. Stability of trend classifications across robustness variants (proportion of 25 topics per method).

Analysis Setting	Variant	Cramér's V (topic × Group SINTA)	
		LDA	BERTopic
BERTopic	All documents as denominator (including outliers)	88	96
BERTopic	Excluding IJASEIT	84	80
BERTopic	Stable journal panel (96 journals)	96	96
LDA	Coverage-matched (highest-confidence documents)	92	96
LDA	Excluding IJASEIT	76	96
LDA	Stable journal panel (96 journals)	96	96

Table 10. SINTA groups, original title language, and topic–SINTA associations.

SINTA Group	Number of Journals	Number of Titles	Originally English Titles (%)
S1–S2	20	8,451	91.1
S3	43	5,960	46.9
S4–S6	220	25,121	30.9

DISCUSSION

Each research question is addressed in sequence, distinguishing empirical findings from their interpretation.

RQ1: Which research topics can be identified? Both methods identified similar major themes in Indonesian computer science journals from 2015 to 2024, including information systems, decision support systems, e-learning, the Internet of Things, networking, security, and machine learning applications (Tables 5 and 6). BERTopic's original 121-topic solution revealed more specialized themes, such as blockchain, cryptography, chatbots, and COBIT governance. The corpus also included topics related to finance, accounting, agriculture, and food studies. These findings reflect SINTA indexed journals rather than the entire landscape of computer science research in Indonesia.

RQ2: How do LDA and BERTopic differ? LDA assigned all titles to topics and produced more balanced topic sizes (956–3,953 titles), whereas BERTopic left 38.5% of titles unassigned and identified smaller, more specialized topics (73–3,166 titles). Agreement between the methods was moderate (NMI = 0.31), with only 11 of the 25 LDA topics having strong one-to-one counterparts in BERTopic. LDA's emerging topics tended to emphasize methodological terms, while BERTopic distributed their titles across application-oriented topics. This difference may reflect how the two methods represent text, although this explanation was not directly tested. Removing 17 generic terms increased diversity by 0.113 without improving coherence, highlighting the challenges of topic modeling with short academic titles.

RQ3: How do LDA and BERTopic compare in topic quality, and what are the implications of their differences for identifying research trends from short academic titles? BERTopic achieved higher coherence ((Cv=0.671)) and

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

topic diversity (0.841) than LDA ((Cv=0.580)); diversity = 0.693) at 25 topics. However, BERTopic assigned topics to only 61.5% of titles, whereas LDA covered the entire corpus. Its higher scores therefore do not establish overall superiority. Although BERTopic produced more specific top terms, some topics remained heterogeneous after hierarchical reduction, and interpretability was not evaluated by human raters.

The two methods offer complementary strengths for trend analysis. LDA provides complete document coverage, while BERTopic identifies more specialized topics that may reveal emerging research areas. Their agreement on several trends, including sentiment analysis, provides additional support for those findings. However, the limited agreement between their topic assignments and the sensitivity of some trends to journal composition, SINTA rankings, and original language highlight the need for cautious interpretation. Using both methods and prioritizing trends that remain robust across statistical tests and corpus-composition checks provides a more defensible approach than selecting one method as universally superior.

RQ4: How did topics develop from 2015 to 2025? Tujuh Seven method-topic combinations remained supported across all robustness checks. These included increases in sentiment analysis, deep learning and convolutional neural networks, the Internet of Things and monitoring systems, medical image analysis, and swarm intelligence and applied machine learning, as well as a decline in learning media and e-learning. Two additional trends, increasing data mining and declining augmented reality and educational games, showed consistent directions but did not retain the same classifications across all tests. COVID-19 topics peaked in 2021–2022 before declining, whereas chatbot and blockchain topics grew from relatively small bases. These findings describe patterns in the study corpus during 2015–2024 rather than the entire landscape of computer science research in Indonesia.

The robustness checks also revealed that some apparent trends were sensitive to corpus composition. The steep declines in Electrical Energy and Power Systems under LDA and Plant Image Classification and Agricultural/Food Studies under BERTopic did not persist after IJASEIT was excluded or after SINTA standardization. Trends in Web-Based Information Systems and Machine Learning and Intelligent Detection were also affected by changes in SINTA composition. Because the representation of predominantly English language and multidisciplinary S1 journals changed over time, some observed trends may reflect shifts in the journal composition of the corpus rather than changes in research activity.

5. Conclusion

This study compared LDA and BERTopic for identifying research topics and examining their temporal development in 39,532 article titles from 283 Indonesian SINTA-indexed computer science journals published between 2015 and 2025. The findings reveal a trade-off between topic quality and document coverage. BERTopic achieved higher coherence and diversity, whereas LDA assigned all titles to topics and produced a more balanced topic-size distribution. Although both methods identified emerging, declining, and persistent research themes, their temporal classifications did not always agree. These differences demonstrate that the choice of topic modeling method can influence how research trends are identified and interpreted. The robustness analysis further showed that some apparent trends were sensitive to journal concentration, SINTA accreditation composition, and the original language of article titles. Accordingly, research trend analysis should evaluate topic quality alongside document coverage, statistical consistency, and corpus composition rather than rely on a single model or trend measure. This study is limited by its reliance on article titles, incomplete data for 2025, and the absence of formal human evaluation of topic interpretability. Future studies could incorporate abstracts or full-text articles, conduct systematic human evaluations, and examine other disciplines and publication databases to assess the generalizability of these findings.

References

[1] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(suppl. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101
[2] Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In Proceedings of the 22nd International Conference on World Wide Web (WWW 2013) (pp. 1445–1456). ACM. https://doi.org/10.1145/2488388.2488514
[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.
[4] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794
[5] Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Frontiers in Sociology, 7, 886498. https://doi.org/10.3389/fsoc.2022.886498
[6] Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics, 8, 439–453. https://doi.org/10.1162/tacl_a_00325

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

[7] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT networks. In Proceedings of EMNLP-IJCNLP 2019 (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410
[8] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426. https://doi.org/10.48550/arXiv.1802.03426
[9] Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Advances in Knowledge Discovery and Data Mining (PAKDD 2013), LNCS 7819 (pp. 160–172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14
[10] McInnes, L., Healy, J., Astels, S. (2017). hdbscan: Hierarchical density based clustering. Journal of Open Source Software, 205. https://doi.org/10.21105/joss.00205 [11] Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In Proceedings of EMNLP 2020 (pp. 4512–4525). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.365
[12] Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM 2015) (pp. 399–408). ACM. https://doi.org/10.1145/2684822.2685324
[13] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
[14] Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (pp. 45–50). ELRA.
[15] Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193–218. https://doi.org/10.1007/BF01908075
[16] Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research, 11, 2837–2854.
[17] Kendall, M. G. (1938). A new measure of rank correlation. Biometrika, 30(1–2), 81–93. https://doi.org/10.1093/biomet/30.1-2.81
[18] Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association, 63(324), 1379–1389. https://doi.org/10.1080/01621459.1968.10480934
[19] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
[20] Simpson, E. H. (1951). The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society: Series B, 13(2), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x
[21] Cramér, H. (1946). Mathematical Methods of Statistics. Princeton University Press.
[22] Lin, J. (1991). Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 37(1), 145–151. https://doi.org/10.1109/18.61115
[23] Virtanen, P., Gommers, R., Oliphant, T. E., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17, 261–272. https://doi.org/10.1038/s41592-019-0686 2 [24] Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In Proceedings of the 9th Python in Science Conference (pp. 92–96).

DOI: https://doi.org/10.69693/ijmst.v4i3.14485
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)