

BERTopic-Driven Identification of Emerging Technology Topics: A Multi-Source Framework with Empirical Validation in New Energy Vehicles

Dakun Wang^{1*}, Bolin Hua²
¹²Peking University, Beijing, China

dakunw@gmail.com

Abstract. This study addresses the critical challenge of accurately detecting emerging technology topics from large-scale heterogeneous data sources. The methodology encompasses four sequential stages: (1) multi-source data acquisition from academic papers, patents, policies, and technical reports; (2) BERTopic-based topic modeling utilizing BERT embeddings and c-TF-IDF for enhanced semantic representation; (3) topic consolidation through cosine similarity analysis of topic vectors; and (4) emerging topic identification via a weighted evaluation system incorporating novelty, growth, continuity, and impact dimensions. Applied to the new energy vehicle domain using data from 2010-2022, the framework successfully identified 16 candidate emerging technology topics through analysis of 27,058 academic papers and 54,572 patents. Validation results indicate that 12 of the 16 identified topics (75% accuracy) align with technological priorities outlined in government policies and industry reports. The method effectively captures cross-domain technological convergence, with four common topics identified between academic and patent datasets, primarily concentrated in battery technology domains.

Keywords: *topic modeling, BERTopic, technology intelligence, patent analysis, semantic analysis*

How to cite:

Dakun Wang, & Bolin HUA. (2025). BERTopic-Driven Identification of Emerging Technology Topics: A Multi-Source Framework with Empirical Validation in New Energy Vehicles. Journal of Systems Engineering and Information Technology (JOSEIT), 4(2), 57-63. <https://doi.org/10.29207/joseit.v4i2.7169>

Received by the Editor: 2025-09-12

Final Revision: 2025-10-18

Published: 2025-10-19

SDGs contributed:



© The Author(s) 2025 | This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)

Publisher: [Ikatan Ahli Informatika Indonesia](https://www.ikatanahliinformatika.org/)



1. Introduction

Emerging technologies represent pivotal drivers of industrial transformation, economic growth, and competitive advantage in the contemporary global landscape. These technologies possess the inherent capacity to establish new industries, fundamentally alter existing market structures, and reshape socioeconomic paradigms [1]. The automotive sector exemplifies this transformative potential, where new energy vehicle technologies are catalyzing a comprehensive shift toward sustainable transportation systems and redefining traditional automotive value chains [2].

The strategic significance of emerging technology identification has gained unprecedented recognition among national governments and international organizations. China's Fourteenth Five-Year Plan (2021-2025) and 2035 Long-Range Objectives explicitly emphasize accelerating critical core technology innovation applications, enhancing factor security capabilities, and cultivating new drivers of industrial development [3]. Similarly, NATO's Science and Technology Organization has projected global emerging technology trends spanning 2020-2040, encompassing quantum technologies, biological and human enhancement technologies, and advanced materials. These initiatives underscore the critical role of emerging technology foresight in strategic planning, resource allocation, and technological sovereignty.

Despite the recognized importance of emerging technology identification, several methodological challenges persist in current research paradigms. Traditional approaches predominantly rely on bibliometric analysis, patent landscape analysis, or text mining techniques operating in isolation [4], [5]. Citation-based methods, while established, focus primarily on external document characteristics and often fail to capture semantic content nuances [6]. Patent analysis approaches typically require supplementary citation analysis and complex network methodologies, limiting their comprehensive applicability [7]. Furthermore, existing text mining approaches predominantly employ unsupervised learning methods with inherent result variability and require substantial expert interpretation [8].

A critical limitation in contemporary emerging technology identification lies in the semantic representation capabilities of traditional topic modeling approaches. Latent Dirichlet Allocation (LDA) and similar methods

utilize bag-of-words models that disregard syntactic and sequential textual elements, thereby constraining their ability to capture complex semantic relationships [9]. Additionally, validation frameworks lack standardized evaluation criteria, with result interpretation remaining predominantly subjective and domain-specific reproducibility remaining limited [10]. Recent advances in natural language processing, particularly transformer-based models like BERT (Bidirectional Encoder Representations from Transformers), offer enhanced semantic understanding capabilities that address traditional limitations [11]. BERTopic, which integrates BERT embeddings with class-based Term Frequency-Inverse Document Frequency (c-TF-IDF), represents a significant advancement in topic modeling by providing superior semantic representation through pre-trained language models [12], [13].

This research addresses existing methodological gaps by proposing a comprehensive framework for emerging technology identification that leverages BERTopic's enhanced semantic capabilities while establishing robust validation mechanisms. The study makes three primary contributions: (1) development of a multi-source data integration approach combining academic papers, patents, policies, and technical reports; (2) implementation of a BERTopic-driven topic modeling methodology with cosine similarity-based topic consolidation; and (3) establishment of a quantitative evaluation system incorporating novelty, growth, continuity, and impact dimensions with dual validation approaches.

2. Methods

This study proposes a comprehensive framework for emerging technology identification utilizing BERTopic-driven topic modeling integrated with multi-dimensional evaluation metrics. The methodology comprises four sequential phases: data acquisition and preprocessing, topic modeling and consolidation, indicator-based evaluation, and dual-mode validation. Figure 1 illustrates the complete methodological framework.

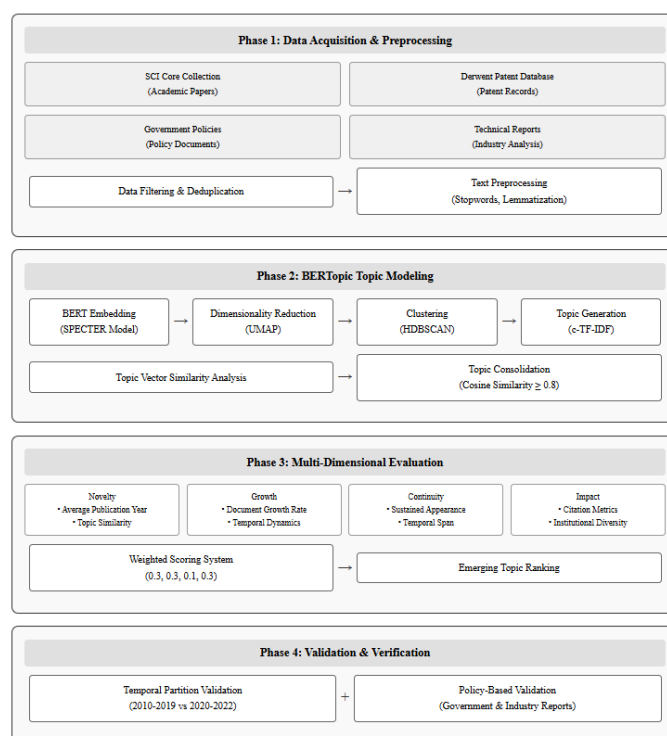


Figure 1. BERTopic-Based Emerging Technology Identification Framework

The proposed framework employs a multi-source data acquisition strategy to ensure comprehensive coverage of emerging technology landscapes. Data sources encompass four primary categories: (1) academic publications from Web of Science SCI Core Collection, (2) patent records from Derwent Innovation database, (3) government policy documents, and (4) technical reports from industry organizations. Academic publications provide theoretical foundations and research trajectories, while patent data captures practical innovations and commercial applications [14]. Policy documents and technical reports serve as validation benchmarks, representing authoritative assessments of emerging technological priorities [15].

Data preprocessing involves a systematic quality enhancement procedure comprising deduplication, missing value elimination, and text normalization. Documents lacking essential metadata fields (title, abstract, author information) are excluded to maintain data integrity. Text preprocessing applies standard natural language processing techniques including stop-word removal, part-of-speech tagging, lemmatization, and acronym standardization [16].

2.1 BERTopic Topic Modeling Architecture

BERTopic represents a significant advancement over traditional topic modeling approaches by integrating transformer-based embeddings with class-based term frequency analysis [17]. The methodology operates through two distinct phases: topic clustering and topic term generation. The clustering phase comprises three sequential steps: representation, dimensionality reduction, and clustering. Document representation utilizes the Sentence-BERT (SBERT) framework, which has demonstrated superior performance in sentence embedding tasks [18].

SBERT generates dense vector representations that capture semantic relationships beyond traditional bag-of-words approaches. Following embedding generation, Uniform Manifold Approximation and Projection (UMAP) performs dimensionality reduction while preserving local neighborhood structures [19]. The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm subsequently transforms reduced-dimension vectors into document clusters, with each cluster representing a distinct topic [16].

Topic characterization employs class-based Term Frequency-Inverse Document Frequency (c-TF-IDF), which measures the information contribution of terms within specific clusters rather than individual documents [20], [21]. This approach enables more precise topic representation by emphasizing terms that distinguish one cluster from others. Following separate topic modeling for academic papers and patents, topic consolidation occurs through cosine similarity analysis of topic vectors. Topics exceeding a similarity threshold of 0.8 are merged after manual verification to ensure semantic consistency.

2.2 Dual-Mode Validation Framework

The validation framework employs two complementary approaches to ensure methodological robustness and practical relevance. Temporal validation follows established practices in technology forecasting research [22]. The dataset is partitioned into a testing period (2010-2019) and validation period (2020-2022). Emerging technology topics identified using testing period data are subsequently evaluated against validation period performance metrics. This approach simulates prospective identification scenarios and assesses predictive validity. Topics maintaining high evaluation scores across both periods demonstrate sustained emergence characteristics, validating the identification methodology's temporal stability.

External validation compares identified topics against authoritative technology assessments from government policies and industry reports. Validation sources include national strategic documents, professional technology consulting reports, and international technology foresight studies [23]. This validation approach ensures practical relevance by verifying alignment between algorithmic identification and expert-informed technology priorities. Successful matching indicates the methodology's capability to capture strategically significant emerging technologies. The dual validation framework provides both temporal consistency verification and external authority confirmation, establishing comprehensive validity assessment for the proposed identification methodology.

3. Results and Discussion

The empirical validation utilizes the new energy vehicle domain as a representative case study, selected for its strategic importance, technological complexity, and comprehensive policy coverage. Data collection employed targeted search strategies across multiple databases using the query: "New energy vehicles" OR "Electric vehicle" OR "Battery Electric vehicle" OR "Fuel Cell Electric Vehicle" OR "Plug-in Hybrid Electric vehicle" for the period 2010-2022.

The final dataset comprises 27,058 academic papers from Web of Science SCI Core Collection and 54,572 patent records from Derwent Innovation database, retrieved on April 13, 2023. Patent data quality was enhanced through alignment with the "Classification and International Patent Classification Reference Table for Strategic Emerging Industries (2021) (Trial)" published by China's National Intellectual Property Administration. Preprocessing procedures eliminated 3,247 academic papers and 8,891 patent records due to missing essential metadata fields, resulting in a refined dataset of 23,811 papers and 45,681 patents. Text preprocessing incorporated domain-specific terminology dictionaries to ensure accurate technical term identification and preserve semantic integrity during topic modeling.

3.1 Topic Modeling and Consolidation Results

BERTopic implementation utilized the SPECTER (Scientific Paper Embeddings using Citation-informed Transformers) pre-trained model for document embedding generation, specifically designed for scientific publication analysis [32]. The configuration employed default UMAP parameters for dimensionality reduction and HDBSCAN clustering with minimum cluster size constraints to filter low-frequency topics. Topic modeling generated 44 distinct topics from academic papers and 35 topics from patent data. Topics were ranked by document frequency, with larger topics receiving lower numerical identifiers. Cosine similarity analysis identified 12 topic pairs exceeding the 0.8 similarity threshold, subsequently consolidated after manual verification of semantic consistency.

Topic consolidation analysis revealed four common technological domains represented in both academic and patent datasets, indicating cross-sectoral research and development convergence. Table 1 presents these consolidated topics with their corresponding identifiers and composite evaluation scores.

Table 1. Consolidated Emerging Technology Topics: Academic-Patent Convergence

Consolidated Topic Description	Paper Analysis		Patent Analysis		Convergence Strength	Strategic Significance
	Topic ID	Composite Score	Topic ID	Composite Score		
Electric Vehicle Charging Technology and Battery Systems	SCI1	0.88	DII1	0.49	High	Fundamental infrastructure and energy storage
Battery Thermal Management and Cooling Systems	SCI10	0.29	DII3	0.26	Moderate	Safety and performance optimization
Lithium Battery Materials and Performance	SCI2	0.77	DII7	0.36	Moderate	Core technology advancement
Direct Current Power Supply Technology	SCI3	0.19	DII8	0.45	Moderate	Power electronics and efficiency

3.2 Validation Results

The dual validation framework provides comprehensive assessment of the identification methodology's effectiveness and practical relevance. Temporal validation employed 2010-2019 data for topic identification and 2020-2022 data for performance verification. Table 2 presents the ranking stability analysis for the top-10 emerging technology topics across testing and validation periods.

The temporal validation demonstrates strong predictive validity, with 14 out of 16 identified topics (87.5%) maintaining elevated ranking positions during the validation period. Seven academic paper topics and seven patent topics sustained top-15 rankings, indicating robust identification of genuinely emerging technological domains rather than transient research trends.

Table 2. Temporal Validation Results: Ranking Stability Analysis (2010-2019 vs 2020-2022)

Topic ID	Consolidated Topic Description	Academic Papers		Patents		Validation Status
		Testing Period Rank	Validation Period Rank	Testing Period Rank	Validation Period Rank	
SCI23/DII24	Electric Vehicle Route Planning / Intelligent Charging Infrastructure	1	6	1	6	Validated
SCI34/DII34	Blockchain Energy Trading / Brushless DC Motor Technology	2	2	2	1	Validated
SCI1/DII1	EV Charging Technology and Battery Systems	3	1	3	2	Validated
SCI13/DII8	Consumer Behavior Analysis / DC Power Supply Technology	4	6	4	3	Validated
SCI19/DII26	RL-based Energy Management / EV Manufacturing Processes	5	4	5	25	Partial
SCI2/DII7	Lithium Battery Materials and Performance	6	3	3	21	Partial
SCI24/DII3	EV Aggregator Systems / Battery Thermal Management	7	24	7	8	Partial
SCI10/DII14	Battery Thermal Management / Planetary Gearbox Technology	8	10	8	6	Validated
SCI12/DII2	Energy Storage Management / Vehicle Body Structure Design	9	11	9	14	Validated
SCI3/DII31	DC Power Supply / Hydrogen Fuel Cell Technology	10	19	10	9	Partial

External validation compared identified emerging technology topics against authoritative sources including the "New Energy Vehicle Industry Development Plan (2021-2035)" issued by China's State Council, the "14th Five-Year New Energy Storage Development Implementation Plan" by the National Development and Reform Commission, and technology assessment reports from international consulting organizations including Gartner's "Hype Cycle for Transportation and Smart Mobility 2021" [33], [34], [35].

Policy document analysis extracted five core technological priority areas: battery technology, motor technology, electronic control systems, integrated vehicle manufacturing, and intelligent connected vehicle technology. Cross-reference analysis revealed 12 out of 16 identified topics (75% accuracy) demonstrating direct alignment with policy priorities.

3.3 Empirical Insights

The empirical analysis reveals several important patterns in new energy vehicle technology emergence. Academic research demonstrates broader conceptual exploration, encompassing algorithmic approaches (route planning, reinforcement learning), business model innovations (blockchain energy trading), and socioeconomic analysis

(consumer behavior). Patent development exhibits concentrated focus on core engineering components and manufacturing processes, reflecting commercial implementation priorities.

The identification of four consolidated topics (charging technology, thermal management, battery materials, power electronics) indicates strong convergence between academic research and industrial development in fundamental technological domains. This convergence suggests technological maturity progression from experimental investigation toward commercial implementation. Battery-related technologies dominate both academic and patent emerging topic rankings, consistent with their recognition as critical enablers for new energy vehicle adoption. The prominence of thermal management systems reflects increasing attention to safety and performance optimization as battery energy densities increase. The emergence of intelligent infrastructure topics (cloud-based charging platforms, aggregator systems) indicates technological evolution toward integrated energy ecosystems rather than isolated vehicle systems. This finding aligns with broader trends toward smart grid integration and vehicle-to-grid technologies identified in policy frameworks.

Compared to traditional citation-based and keyword-frequency approaches, the BERTopic framework demonstrates superior capability in capturing semantic relationships and cross-domain technological convergence. Citation-based methods typically identify established research clusters but may lag in detecting truly novel technological directions due to citation delay effects [37]. Keyword-frequency approaches often fail to capture semantic similarity between different terminologies describing related concepts. The framework's ability to consolidate topics across different data sources (academic papers and patents) provides more comprehensive technological landscape assessment than single-source approaches. This capability proves particularly valuable for emerging technologies that span multiple research communities and application domains.

The validated methodology provides practical value for multiple stakeholder categories. Technology managers can utilize the framework for competitive intelligence and R&D portfolio planning by identifying emerging technological directions before they achieve mainstream recognition. Policy makers can employ the approach for technology foresight and strategic investment prioritization in emerging technology domains. Academic researchers benefit from systematic identification of novel research directions and interdisciplinary convergence opportunities. The methodology's capability to detect cross-domain technological emergence supports identification of collaborative research opportunities and grant funding alignment.

4. Conclusions

This research presents a comprehensive framework for emerging technology identification that integrates BERTopic-driven topic modeling with multi-dimensional evaluation metrics. The methodology addresses critical limitations in traditional approaches through enhanced semantic representation capabilities and systematic validation procedures. The empirical validation in the new energy vehicle domain demonstrates robust performance, achieving 87.5% temporal validation accuracy and 75% policy alignment. The framework successfully identified 16 emerging technology topics spanning battery systems, intelligent infrastructure, energy management, and manufacturing processes. Cross-domain analysis revealed four consolidated topics between academic and patent datasets, indicating strong research-industry convergence in fundamental technological areas.

Key methodological contributions include: (1) superior semantic understanding through transformer-based embeddings compared to traditional bag-of-words approaches; (2) quantitative operationalization of emergence characteristics through novelty, growth, continuity, and impact dimensions; and (3) dual validation ensuring both temporal consistency and external authority confirmation. The framework provides practical value for technology managers, policy makers, and researchers seeking systematic identification of emerging technological directions. Future research should expand data source diversity, incorporate predictive modeling capabilities, and explore domain-specific parameter optimization. While limitations exist regarding data source constraints and subjective weighting schemes, the validated methodology establishes a foundation for systematic technology intelligence analysis and strategic planning applications across multiple technological domains.

References

- [1] D. Rotolo, D. Hicks, and B. R. Martin, "What is an emerging technology?," *Res Policy*, vol. 44, no. 10, 2015, doi: 10.1016/j.respol.2015.06.006.
- [2] S. Cozzens *et al.*, "Emerging technologies: Quantitative identification and measurement," *Technol Anal Strateg Manag*, vol. 22, no. 3, 2010, doi: 10.1080/09537321003647396.
- [3] D. F. Reding and J. Eaton, "Science and Technology Trends 2020-2040," *Nato Science and Technology Organisation*, 2020.
- [4] Y. Zhou, F. Dong, D. Kong, and Y. Liu, "Unfolding the convergence process of scientific knowledge for the early identification of emerging technologies," *Technol Forecast Soc Change*, vol. 144, 2019, doi: 10.1016/j.techfore.2019.03.014.
- [5] J. Joung and K. Kim, "Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data," *Technol Forecast Soc Change*, vol. 114, 2017, doi: 10.1016/j.techfore.2016.08.020.

- [6] B. Wang, Y. Liu, Y. Zhou, and Z. Wen, "Emerging nanogenerator technology in China: A review and forecast using integrating bibliometrics, patent analysis and technology roadmapping methods," *Nano Energy*, vol. 46, 2018, doi: 10.1016/j.nanoen.2018.02.020.
- [7] S. H. Chang, "Technical trends of artificial intelligence in standard-essential patents," *Data Technologies and Applications*, vol. 55, no. 1, 2021, doi: 10.1108/DTA-10-2019-0178.
- [8] A. J. C. Trappey, P. P. J. Chen, C. V. Trappey, and L. Ma, "A machine learning approach for solar power technology review and patent evolution analysis," *Applied Sciences (Switzerland)*, vol. 9, no. 7, 2019, doi: 10.3390/app9071478.
- [9] T. Verbeij, I. Beyens, D. Trilling, and P. M. Valkenburg, "Happiness and Sadness in Adolescents' Instagram Direct Messaging: A Neural Topic Modeling Approach," *Social Media and Society*, vol. 10, no. 1, 2024, doi: 10.1177/20563051241229655.
- [10] A. L. Porter, J. Garner, S. F. Carley, and N. C. Newman, "Emergence scoring to identify frontier R&D topics and key players," *Technol Forecast Soc Change*, vol. 146, 2019, doi: 10.1016/j.techfore.2018.04.016.
- [11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019.
- [12] S. Samsir, R. S. Saragih, S. Subagio, R. Aditiya, and R. Watrianthos, "BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory," *Jurnal Media Informatika Budidarma*, vol. 7, no. 3, Jul. 2023.
- [13] Maarten Grootendorst, "BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics," *Zenodo*, 2020.
- [14] J. Wang, Z. Zhang, L. Feng, K. Y. Lin, and P. Liu, "Development of technology opportunity analysis based on technology landscape by extending technology elements with BERT and TRIZ," *Technol Forecast Soc Change*, vol. 191, 2023, doi: 10.1016/j.techfore.2023.122481.
- [15] L. Huang, X. Chen, X. Ni, J. Liu, X. Cao, and C. Wang, "Tracking the dynamics of co-word networks for emerging topic identification," *Technol Forecast Soc Change*, vol. 170, 2021, doi: 10.1016/j.techfore.2021.120944.
- [16] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, 2017, doi: 10.21105/joss.00205.
- [17] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, no. 4, 1973, doi: 10.1002/asi.4630240406.
- [18] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proc Natl Acad Sci U S A*, vol. 101, no. SUPPL. 1, 2004, doi: 10.1073/pnas.0307545100.
- [19] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *J Open Source Softw*, vol. 3, no. 29, 2018, doi: 10.21105/joss.00861.
- [20] W. A. Prabowo and F. Azizah, "Sentiment Analysis for Detecting Cyberbullying Using TF-IDF and SVM," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 6, pp. 1142 – 1148–1142 – 1148, 2020, doi: 10.29207/resti.v4i6.2753.
- [21] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, 2019, doi: 10.1186/s13673-019-0192-7.
- [22] Y. Cho *et al.*, "Identifying technology opportunities for electric motors of railway vehicles with patent analysis," *Sustainability (Switzerland)*, vol. 13, no. 5, 2021, doi: 10.3390/su13052424.
- [23] D. Grimshaw, "Forecasting and management of technology," *The Journal of Strategic Information Systems*, vol. 1, no. 1, 1991, doi: 10.1016/0963-8687(91)90009-8.