Hybrid Intelligence Framework for Identifying Frontier Technologies through Project Linkage: A Case Study of DARPA Programs

Xiang-li Zhu, Yifan Wang, Xiaoping Liu

National Science Library, Chinese Academy of Sciences; School of Economics and Management, University of Chinese Academy of Sciences

Beijing, China

email: zhuxl@mail.las.ac.cn

***Abstract*—In the context of the global technological revolution and industrial transformation, the identification of frontier technologies has become a critical component of national strategic competition. However, traditional methods based on citation analysis or patent classification suffer from significant time lags and fail to comprehensively capture the entire lifecycle from technological conception to practical application. To address this, this paper proposes a novel project-linkage paradigm for frontier technology identification, constructing an integrated framework that combines data-driven analysis, intelligent algorithms, and multidimensional assessment. The framework utilizes Large Language Models (LLMs, such as DeepSeek V3) to enhance textual feature extraction and combines Word2Vec vectorization with K-means clustering for technical topic discovery, establishing technology evolution chains through cross-source semantic associations between project requirements and research outputs. Using Defense Advanced Research Projects Agency (DARPA)-funded programs from 2009 to 2025 as empirical subjects, the study finds that: (1) the response rate of research projects to academic publications increased significantly from 80% to 98.3%, indicating that DARPA projects are shifting from following academia to leading academia; (2) three tiers of frontier technologies were identified—mature frontiers, emerging frontiers, and potential frontiers. The results show that the proposed hybrid intelligence framework effectively identifies prospective technological breakthroughs, offering precise support for science and technology decision-making.**

Keywords—Hybrid Intelligence; Frontier Technology Identification; Multi-source Data Fusion; Research Projects; Large Language Models.

I. INTRODUCTION

In the global competition for technological leadership, frontier technologies serve as strategic assets, and their early identification is crucial for securing a competitive edge. However, traditional methods of technology identification, such as citation analysis and patent classification, rely on single data sources and suffer from significant time lags, making it difficult to capture the full trajectory from fundamental research to applied innovation [1][2]. While research projects are the primary point of resource allocation in science and technology, they encapsulate explicit strategic orientations and forward-looking requirements, often containing valuable insights into future technological directions. Previous studies have largely treated project metadata as secondary information, overlooking the rich technological signals embedded within project texts.

Recent advancements in big data analytics and artificial intelligence, particularly the capabilities of LLMs for semantic understanding, have opened new possibilities for identifying frontier technologies from large and diverse datasets [3]. This underscores the need for novel methodologies that integrate multi-source data—such as project descriptions, academic publications, and patents—along with intelligent algorithms to achieve more timely, accurate, and comprehensive identification of frontier technologies.

To address this gap, this study introduces a new project-linkage paradigm for frontier technology identification. Unlike traditional methods, we position research projects as the starting point, conducting semantic association analysis between project descriptions and their subsequent publications and patents. This enables cross-source mapping, connecting project inputs (funding and requirements) to research outputs (publications and patents). We propose a data-driven framework that leverages Large Language Models to enhance text extraction capabilities, integrates vectorization and clustering techniques to ensure reproducibility, and incorporates a multidimensional assessment system to classify identified technologies by maturity. To validate our methodology, we conduct an empirical study using DARPA-funded projects and their associated outputs from 2009 to 2025, analyzing the linkage between projects and publications/patents to identify emerging frontier technologies and categorize them based on their development stage.

The main contributions of this work are threefold: (1) A novel project-linkage paradigm that positions research project texts as primary signals for frontier technology identification, shifting from traditional publication/patent-centric approaches; (2) A hybrid intelligence pipeline that synergistically combines LLM-based semantic extraction with reproducible vectorization-clustering workflows, enabling dual-verification through both computational similarity and LLM-based relevance scoring; (3) A five-dimensional maturity assessment system that stratifies identified technologies into mature, emerging, and potential frontiers, providing actionable intelligence for decision-makers.

The remainder of this paper is organized as follows. In Section II, we review the related work on frontier technology conceptualization and identification methods, including recent advances in AI-driven approaches. In Section III, we present the proposed hybrid intelligence framework, covering technical phrase extraction, semantic vectorization, cross-

 ISBN: 978-1-68558-397-2

source association analysis, and the multidimensional assessment system. In Section IV, we describe the empirical study using DARPA unmanned systems projects and present the results. In Section V, we conclude the paper and discuss future research directions.

II. Related Work

This section reviews the literature in three areas: the conceptual evolution of frontier technology, methods for its identification, and recent AI-driven advances that underpin our approach.

A. *Evolution of Frontier Technology Conceptualization*

The concept of frontier technology lacks a universal definition, with different scholars offering varying interpretations. In 1965, Price [4] introduced the concept of research fronts, defining them as research areas represented by a set of recently published and frequently cited papers, emphasizing novelty and academic attention. In the 21st century, the focus shifted to technological characteristics and industrial value: Chen [5] noted that an abrupt increase in technology occurrence frequency signals the emergence of frontier technologies; Cozzens et al. [6] argued that frontier technologies should feature rapid growth, novelty, untapped market potential, and strong technological foundations; Rotolo et al. [7] identified five key attributes of frontier technologies: novelty, rapid growth, coherence, potential impact, and uncertainty/ambiguity.

In policy and governance contexts, international organizations generally view frontier technologies as rapidly developing technologies with high uncertainty, capable of generating significant socioeconomic impacts. The United Nations Conference on Trade and Development (UNCTAD) Technology and Innovation Report [8] defines frontier technologies as new and rapidly developing technologies, emphasizing their economic potential and technological gaps. The World Intellectual Property Organization (WIPO) highlights that frontier technologies lie at the intersection of scientific breakthroughs and real-world applications, while the Organisation for Economic Co-operation and Development (OECD) underscores the dual nature of emerging technologies and the need for forward-thinking policies and risk governance [9].

B. *Frontier Technology Identification Methods*

As the concept of frontier technology has evolved, so too have methods for identifying them, transitioning from single-source to multi-source, static to dynamic, and manual to automated approaches. Early methods mainly relied on bibliometric techniques, such as citation networks, co-word analysis, and keyword burst detection to reveal research frontiers. However, these methods often lag behind actual technological development. Subsequently, patent data was introduced to track technological innovations, but publications and patents have inherent limitations: publications highlight scientific novelty but may lack industrial relevance, while patents reflect applied innovations with limited coverage.

Recently, multi-source data fusion has gained traction, combining various data sources to overcome the shortcomings of single-source methods. Liu et al.[10] proposed integrating publications, patents, startup data, and public opinions to create a multidimensional indicator system for identifying disruptive technologies, proving that multi-source fusion outperforms single-source approaches. Munari et al. [11] introduced a research project/funding-publication-patent linkage method, demonstrating how public funding influences technology output and providing a framework for earlier frontier identification using project data.

C. *AI-Driven and Multi-Source Identification Methods*

Recent advances in Natural Language Processing (NLP) and deep learning have introduced powerful tools for technology identification. In the domain of semantic embedding, Mikolov et al. [12] proposed Word2Vec, which learns distributed word representations that capture syntactic and semantic relationships. More recently, Devlin et al. [10] introduced Bidirectional Encoder Representations from Transformers (BERT), enabling contextualized embeddings that significantly improve downstream NLP tasks. Beltagy et al. [14] further developed SciBERT, a variant pre-trained on scientific text, demonstrating superior performance on scientific document processing tasks.

In the area of LLM-assisted topic extraction, Grootendorst [15] proposed BERTopic, a topic modeling technique that leverages transformer-based embeddings and class-based Term Frequency-Inverse Document Frequency (TF-IDF) to create dense topic clusters. More recently, Xu et al. [16] demonstrated that LLMs can be effectively used for automated information extraction from scientific literature, outperforming traditional rule-based and statistical methods. For technology forecasting and emerging technology detection, Huang et al. [17] developed a framework for technology opportunity analysis using multi-source patent data and semantic analysis. These prior works provide methodological foundations upon which our hybrid intelligence framework builds, while our approach distinctively integrates project-level data as early signals and employs a dual-verification mechanism combining computational similarity with LLM-based relevance scoring.

III. Methodology

This section presents the proposed hybrid intelligence framework, covering its overall design, the technical phrase extraction and vectorization pipeline, the project-output association analysis, and the multidimensional assessment system for frontier stratification.

A. *Overall Framework Design*

The proposed project-linkage paradigm for frontier technology identification, as illustrated in Figure 1, is structured into three key stages. First, Technical Topic Extraction & Clustering, involves aggregating project descriptions, publications, and patents from various data sources, followed by a multi-stream processing approach. This utilizes techniques, such as K-means clustering and LLMs for intelligent summarization and topic extraction, resulting in standardized project, publication, and patent topics. Second, cross-Source Semantic Association, integrates the extracted topics through Word2Vec vectorization and cosine similarity, establishing semantic connections across data sources. Subjective association scoring based on LLMs is employed for dual-verification of the similarity calculations, with the results organized and filtered through time-series analysis. The final stage, Multidimensional Assessment & Stratification, evaluates the identified frontier technologies using a five-dimensional indicator system, comprising Response Scale, Semantic Similarity, Temporal Directionality, Evolution Continuity, and Cross-domain Integration. This stage leverages LLMs to provide a comprehensive maturity assessment, categorizing technologies as mature, emerging, or potential frontiers.

B. Technical Phrase Extraction with Large Language Models

Traditional keyword extraction methods based on TF-IDF or TextRank often fail to capture domain-specific technical semantics. This study employs the DeepSeek V3 Large Language Model for technical phrase extraction. The model processes project descriptions and generates concise technical phrases that represent core technological concepts. Specifically, the prompt instructs the model as follows: "*Given the following project description, extract up to 10 concise technical phrases that represent the core technological concepts. Each phrase should be 2–5 words and capture a distinct technical aspect*." This approach leverages the semantic understanding capabilities of LLMs to identify meaningful technical terms that may not be captured by frequency-based methods.

C. Semantic Vectorization and Topic Clustering

Following phrase extraction, we employ Word2Vec for semantic vectorization, converting technical phrases into 300-dimensional vector representations. The K-means clustering algorithm is then applied to group semantically related phrases into distinct technical topics. The optimal number of clusters is determined through a grid search over K = 5 to 30, selecting the value that maximizes the silhouette coefficient, followed by domain expert validation. This dual-layer approach—LLM extraction followed by traditional vectorization and clustering—ensures both semantic depth and computational reproducibility.

D. Project-Output Association Analysis

To establish associations between project topics and research outputs (publications and patents), we calculate cosine similarity between vectorized project topics and output topics. A cosine similarity threshold of 0.15 is applied to filter meaningful associations. Additionally, the LLM generates relevance scores for each project-output pair on a scale from 0 to 100, with a threshold of 60 for qualifying associations. The final association score is computed as: $Final\_Score = \alpha \times CosSim\_norm + (1-\alpha) \times LLM\_Score\_norm$, *where* $\alpha = 0.5$. High-scoring pairs are identified as meaningful associations, forming project-output networks.

E. Multidimensional Assessment System

To evaluate the frontier attributes of identified technologies, we propose a five-dimensional indicator system, as detailed in Table I. The system comprises: Response Scale (A), measuring the volume and proportion of projects yielding relevant outputs; Semantic Similarity (S), quantifying the average topical alignment between projects and their outputs; Temporal Directionality (T), assessing the chronological distribution (prior, concurrent, or posterior) of outputs relative to project initiation; Evolution Continuity (C), tracking the stability of response levels across successive time windows; and Cross-Domain Integration (D), evaluating the coupling degree of topics across heterogeneous domains. By synthesizing these metrics, the paradigm enables a stratified classification of technologies into mature, emerging, and potential frontiers. The framework was implemented in Python 3.10 using Gensim for Word2Vec training, scikit-learn for K-means clustering, and the DeepSeek API for LLM-based extraction. The code will be released as open source upon publication.

IV. EMPIRICAL STUDY

This section describes the empirical validation of the proposed framework using DARPA-funded unmanned systems programs, presents the project-output association analysis results for two temporal phases, and synthesizes the frontier technology directions identified through the multidimensional assessment system.

A. Data and Experimental Setup

This study utilizes data from research projects in unmanned systems technologies funded by the U.S. DARPA. DARPA, a major funder of frontier technology Research and Development (R&D), supports projects in various cutting-edge fields, including advanced unmanned aerial vehicles, unmanned ground vehicles, and robotic swarms. The dataset comprises a collection of DARPA-initiated projects from 2009 to 2025, including project titles, abstracts, technical descriptions, and publicly available outputs (such as publications and patents). These data were primarily sourced from project completion reports, literature databases, and patent databases, ensuring a comprehensive representation of each project lifecycle.

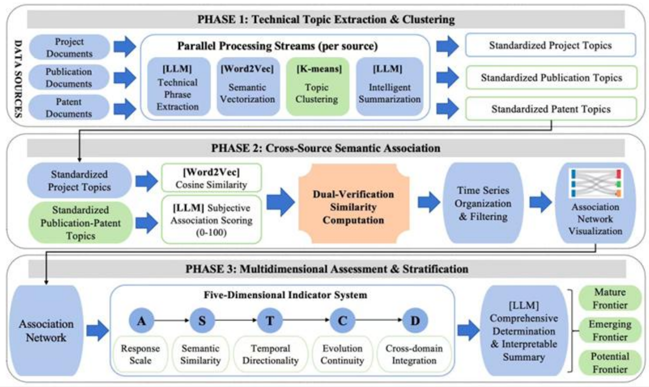

Figure 1. Schematic diagram of the hybrid intelligence framework for project-linkage-based frontier technology identification.

To examine the associations between DARPA-funded projects and their corresponding research outputs, we divide the dataset into two temporal phases: 2014–2019 and 2020–2025. This segmentation allows for a comparative analysis of project-output interactions across different periods of technological development. For technical phrase extraction, the DeepSeek V3 model uses a prompt designed to extract up to 10 concise technical phrases per project. The Word2Vec model utilizes pre-trained 300-dimensional vectors, while the K-means algorithm adjusts the number of clusters based on the silhouette coefficient. The LLM provides relevance scores for project-output topic pairs on a scale from 0 to 100, which are normalized and averaged with cosine similarity results.

B. *Project-Output Association Analysis*

Project-Publication Association. During 2014–2019, 60 project topics and 28 publication topics yielded 166 high-similarity associations with an average similarity of 0.31 (Figure 2(left)), achieving an 80% response rate (48 of 60 projects). In 2020–2025, 60 project topics and 30 publication topics produced 354 associations with a lower average similarity of 0.19 (Figure 2(right)), while the response rate rose sharply to 98.3% (59 of 60 projects). This combination of increased association volume and decreased similarity reflects a semantic diversification of research trajectories rather than weakening ties.

The topological shift between the two periods is significant. The 2014–2019 network exhibits a sparse bipartite structure with isolated clusters, suggesting that project-to-publication spillover was confined to specific technical niches. By contrast, the 2020–2025 network displays markedly higher density and multifocal hubs, with "many-to-many" associations indicating that individual DARPA project topics now catalyze research across multiple academic domains. This structural evolution confirms a transition from a linear technology transfer model to a complex, integrated collaborative network, where DARPA projects act as gravitational centers for increasingly diverse and interdisciplinary research.

Project-Patent Association. During 2014–2019, 60 project topics and 29 patent topics produced 64 high-similarity associations with an average similarity of 0.31 (Figure 3(left)), of which 37 pairs involved post-project patents and 27 leveraged pre-existing patents. The response rate reached 60% (36 projects). In 2020–2025, with a stable topic scale (60 project and 30 patent topics), the average similarity rose significantly to 0.53 (Figure 3(right)), with 74 associations identified—51 involving patents granted during the project lifecycle and 23 utilizing existing intellectual property. However, the response rate moderated to 50%. This divergent pattern—rising similarity but declining response rate—suggests a strategic shift from exploring uncharted technological territories toward high-fidelity application and integration of existing innovations.

TABLE I. ASSESSMENT DIMENSIONS FOR PROJECT-LINKAGE-BASED FRONTIER TECHNOLOGY IDENTIFICATION

Dimension	Indicator	Meaning of Discriminatory Information
Response Scale (*A*)	Number of high-similarity associations between project topics and output topics	Characterizes response scale and coverage of project topics in outputs (diffusion/attention)
Semantic Similarity (*S*)	Average semantic similarity across all valid project topic–output topic associations	Measures semantic alignment between project technical intent and output technical content
Temporal Directionality (*T*)	Proportions of prior/concurrent/posterior outputs relative to project initiation	Determines whether projects drive output production (posterior dominance) or absorb/integrate existing technologies (prior dominance)
Evolution Continuity (*C*)	Sustained response levels across time windows (strong/moderate/weak)	Characterizes whether topics form cross-period continuous evolution chains (short-term hotspot vs. long-term evolution)
Cross-Domain Integration (*D*)	Coupling degree of topics across different domains	Reflects cross-disciplinary and cross-technical characteristics common in frontier technologies

Copyright (c) IARIA, 2026. ISBN: 978-1-68558-397-2

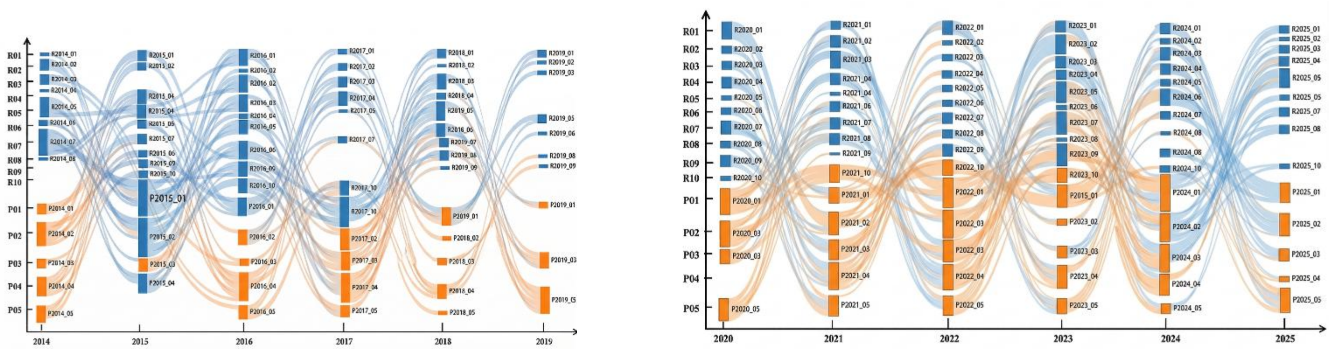

Figure 2. Comparative analysis of high-similarity association networks between DARPA project topics (R) and academic publication topics (P) for the periods 2014–2019 (left) and 2020–2025 (right).

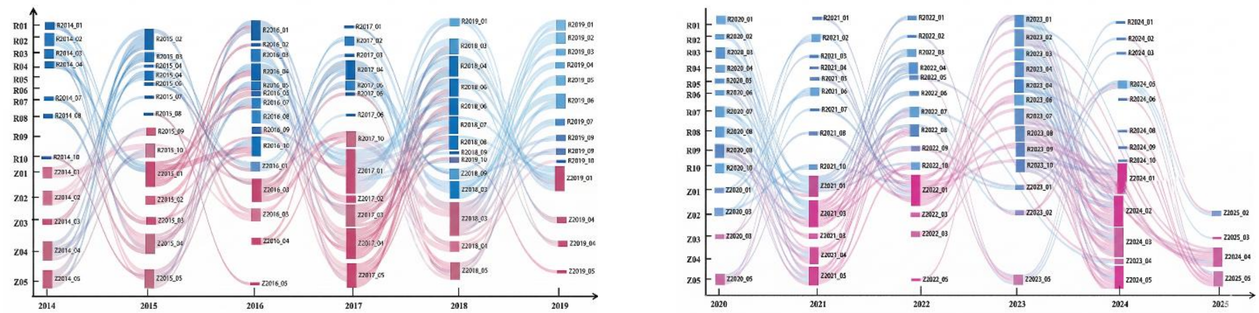

Figure 3. High-similarity association networks between DARPA project topics (R) and patent topics (Z) for 2014–2019 (left) and 2020–2025 (right).

C. Comparative Analysis and Frontier Identification Empowerment

The divergent trajectories of Project-Publication (R-P) and Project-Patent (R-Z) associations provide a high-fidelity lens for characterizing the unmanned systems innovation ecosystem.

Decoupling Research Breadth and Technical Depth: The R-P network exhibits centrifugal expansion (similarity: 0.31→0.19), confirming that DARPA projects catalyze broad, interdisciplinary academic exploration. In contrast, the R-Z network shows centripetal convergence (similarity: 0.31→0.53), suggesting a strategic transition toward high-fidelity application of existing innovations.

Calibrating Technology Readiness Level (TRL) Transitions: The widening "Similarity Gap"—declining R-P similarity coupled with rising R-Z similarity in 2020–2025—marks a critical phase where technologies transition from basic research to system-level integration. The identified "Mature Frontiers" represent directions where both R-P and R-Z associations have stabilized into high-consistency evolution chains.

Strategic Vacuum Detection: The near-saturated publication response (98.3%) coupled with moderated patent response (50%) highlights areas where academic consensus exists but the intellectual property landscape remains fluid, providing precise early-stage breakthrough signals for decision-makers.

Notably, DARPA's strategic priorities shifted around 2019–2020 from platform-centric programs toward multi-domain autonomy and AI-enabled decision-making, partially explaining the observed network densification. While these results are ecosystem-specific, the methodology is domain-agnostic and applicable to other funding agencies (e.g., the National Science Foundation (NSF), European Union (EU) Horizon programs) with appropriate data adaptation.

D. Frontier Direction Identification in Unmanned Systems Technologies

Integrating project-publication and project-patent semantic association results, we conducted frontier attribute assessment of DARPA-deployed unmanned systems project topics based on the five-dimensional indicator system, employing within-sample quantiles (Q25/Q50/Q75) as relative discrimination benchmarks for stratified technology topic identification. Final results, combined with domain expert validation and refinement, yielded a frontier technology direction inventory for unmanned systems. The expert validation involved three domain specialists who independently reviewed and scored the identified frontier directions; a majority voting mechanism was employed to resolve disagreements.

Mature frontier directions primarily exhibit high output response frequency, high semantic consistency, and good evolution continuity. Representative directions include multi-modal sensor fusion chips, autonomous navigation algorithms, and distributed resilient computing platforms.

These technologies have been validated across multiple DARPA projects and formed stable outputs at both publication and patent levels, demonstrating relatively high maturity.

Emerging frontier directions embody strong innovation and cross-domain integration characteristics. While output scale remains limited, growth trends are evident, such as federated learning algorithms, neuromorphic computing chips, and digital twin validation platforms. These directions typically exhibit high semantic consistency and clear project-posterior output features, reflecting rapid exploration phases.

Potential frontier directions are predominantly in early nascent stages, with currently limited output production but clear technological orientation, such as complex terrain autonomous navigation algorithms and platform-independent development middleware. These directions have not yet formed stable evolution chains but possess clear application requirements and subsequent development potential, warranting continuous tracking.

V. CONCLUSION AND FUTURE WORK

Focusing on DARPA-funded project portfolios, this study proposes and validates a project-linkage-based frontier technology identification methodology. We innovatively incorporate project texts into frontier detection perspectives, fusing Large Language Models with bibliometric techniques to realize full-chain association analysis from project requirements to publication and patent outputs. In the empirical study of DARPA unmanned systems projects, this methodology successfully mapped project-output interaction networks, revealed technology diffusion pathways, and consequently identified multiple strategically significant frontier directions.

Research demonstrates that: research project data contains abundant prospective technological intelligence; through intelligent extraction and semantic association, this effectively compensates for deficiencies in single-source publication or patent analysis, significantly enhancing sensitivity to emerging technologies; the integration of multi-source data fusion with LLM technology provides powerful tools for frontier identification, not only automatically generating technical topics and descriptions but also discovering novel cross-domain associations through deep semantic matching, bringing new perspectives to complex technology evolution analysis; introducing multidimensional indicator systems for frontier attribute assessment of technology topics is necessary, ensuring identification results possess greater explanatory power and credibility, facilitating decision-maker comprehension and application.

Naturally, this research has certain limitations. First, regarding data acquisition, heavy reliance on DARPA public materials means some sensitive projects and unpublicized outputs were not included, potentially causing frontier identification omissions. Second, while LLM-generated summaries and similarity scores enhance analysis quality, they also introduce uncertainty and computational overhead, necessitating refined manual verification mechanisms to ensure result reliability. Third, the current thresholds (cosine similarity cutoff of 0.15, LLM relevance score cutoff of 60, fusion weight $\alpha = 0.5$) were determined through manual tuning and expert judgment, rather than automated optimization. Finally, this study's stratification discrimination thresholds and rules may require domain-specific adjustments and are not universally applicable models.

Future work will pursue several directions: systematic comparisons against baseline methods, including pure bibliometric approaches and keyword-based detection methods; ablation studies to quantify the individual contributions of each framework component; sensitivity and robustness analysis across varying thresholds and parameters; computational cost and scalability assessment for larger datasets; automated parameter optimization through grid search, Bayesian optimization, or cross-validation strategies; and cross-domain validation beyond DARPA unmanned systems to assess methodological generalizability across different technological ecosystems and funding agencies.

REFERENCES

[1] W. Glänzel and A. Schubert, "Analysing scientific networks through co-authorship," in *Handbook of quantitative science and technology research*, H. F. Moed, W. Glänzel, and U. Schmoch, Eds. Dordrecht: Kluwer Academic, 2004, pp. 257–276.

[2] A. L. Porter and S. W. Cunningham, *Tech mining: exploiting new technologies for competitive advantage*. Hoboken, NJ: Wiley, 2005.

[3] Y. Guo, J. Huang, H. Chen, and J. Xu, "Identifying technology evolution pathways using topic modeling," J. Informetr., vol. 15, no. 4, Art. no. 101228, 2021.

[4] D. J. de S. Price, "Networks of scientific papers," Science, vol. 149, no. 3683, pp. 510–515, 1965.

[5] C. Chen, "CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature," J. Am. Soc. Inf. Sci. Technol., vol. 57, no. 3, pp. 359–377, 2006.

[6] S. Cozzens, S. Gatchair, J. Kang, K.-S. Kim, H. J. Lee, G. Ordonez, and A. Porter, "Emerging technologies: quantitative identification and measurement," Technol. Anal. Strateg. Manag., vol. 22, no. 3, pp. 361–376, 2010.

[7] D. Rotolo, D. Hicks, and B. R. Martin, "What is an emerging technology?" Res. Policy, vol. 44, no. 10, pp. 1827–1843, 2015.

[8] United Nations Conference on Trade and Development (UNCTAD), *Technology and Innovation Report 2023*. United Nations, 2023.

[9] Organisation for Economic Co-operation and Development (OECD), *Framework for Anticipatory Governance of Emerging Technologies*. OECD Publishing, 2024.

[10] X. Liu, X. Wang, L. Lyu, and Y. Wang, "Identifying disruptive technologies by integrating multi-source data," Scientometrics, vol. 127, no. 9, pp. 5325–5351, 2022.

[11] F. Munari, E. Leonardelli, S. Menini, H. Morais Righi, M. Sobrero, S. Tonelli, and L. Toschi, "Public research funding and science-based innovation: an analysis of ERC research grants, publications and patents," Res. Eval., 2024, doi:10.1093/reseval/rvae012.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in Proc. Int. Conf. Learn. Represent. (ICLR), 2013.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[14] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in Proc. EMNLP-IJCNLP, 2019, pp. 3615–3620.

[15] M. Grootendorst, "BERTopic: neural topic modeling with a class-based TF-IDF procedure," arXiv:2203.05794 [cs.LG], 2022.

[16] S. Xu, L. Shang, and P. Hao, "Automated information extraction from scientific literature using large language models," Scientometrics, vol. 129, no. 4, pp. 2187–2208, 2024.

[17 Y. Huang, L. Zhang, and J. Zhu, "Technology opportunity analysis using multi-source patent data and semantic analysis," Technol. Forecast. Soc. Change, vol. 167, Art. no. 120680, 2021.