Hybrid NLP-Based Knowledge Discovery in Generative AI: A Systematic Literature Review

Shamneesh Sharma
Lincoln Global Postdoctoral & Researcher Programme, Lincoln University College, Petaling Jaya, Selangor Darul Ehsan-47301, Malaysia
(LUC/CPGS/PDF/2026100303)
shamneesh.sharma1987@gmail.com

Shashi Kant Gupta
Lincoln University College, Petaling Jaya, Selangor Darul Ehsan-47301, Malaysia
shashigupta@lincoln.edu.my

Abstract: Research on Generative Artificial Intelligence (GenAI) has accelerated at an unprecedented rate, with an estimated annual growth of 18% in the number of publications across the world making it ever harder for researchers to keep up with new concepts, catch up with new knowledge architectures and predict future research foci. Current knowledge discovery methods mainly utilize statistical topic modelling methods like Latent Dirichlet Allocation (LDA), which provide interpretable topics and limited contextual understanding, or context-sensitive language models like BERT, which provide language model nudity at the cost of lack of interpretability and trend forecasting. This paper presents a novel hybrid approach that combines LDA, LSA and BERT-based contextual embeddings in a PRISMA-guided systematic literature search methodology, covering seven databases. As of now, the literature synthesis reveals a preliminary state of the art of current tools that are fragmented, are good at both statistical coherence and contextual richness, and, most importantly, largely fail to foresee new topics. The proposed framework should enhance the coherence of the topics, facilitate context-aware knowledge discovery, and provide an evidence-based decision-making and technology forecasting automated research-intelligence dashboard in the context of the GenAI space.

Keywords: Generative AI; Natural Language Processing; Topic Modeling; Knowledge Discovery; Systematic Literature Review

Introduction

Growing evidence in recent years shows that GenAI-related publications have grown by almost 18% annually since 1981 [9] and that this is one of the most active research areas in computer science. This is across the growth of large language models, diffusion-based generators and multimodal systems and marked by a quick diversification of research themes. This speed poses a practical problem for individual researchers, research groups and institutions: keeping up with the current themes, which are still developing and which are outmoded can become increasingly problematic, as the literature is scanned manually. Natural Language Processing (NLP) techniques provide scalable solutions for automated knowledge discovery from massive research corpora, converting unstructured data from publications into structured and interpretable knowledge. Current solutions, however, make use of only one family of techniques, either statistical topic models or statistical language models, that can only solve half of the problem. This drives the focus of the central problem tackled in this paper, which is the lack of a consistent, explainable and context-aware NLP framework to map the existing structure of GenAI research and predict its future trajectory. The rest of this paper will be structured as follows. In Section II, related works of statistical and contextual approaches to knowledge discovery are reviewed. The main contribution of

the proposed framework is elaborated on in Section III. The methodology proposed is discussed in Section IV. The implications of it are explored in Section V and the paper concludes in Section VI.

Related work

Scientific literature has been the main context used for the automatic discovery of knowledge so far, which has been done using statistical topic modeling. The most popular model due to the interpretability of the results and ease of computation is the Latent Dirichlet Allocation (LDA) model [1] which represents documents as mixtures of latent topics, each corresponding to a pattern of co-occurrence of words. The earlier matrix factorization method, called Latent Semantic Analysis (LSA) [2] also discovers latent semantic structure with some less success at distinguishing fine-grained topics in large and heterogeneous corpora. In either approach, words are treated as unordered counts, which do not incorporate the notion of context, and this sometimes leads to semantically related but lexically distinct terms being classified as unrelated, which is a common problem in a number of subfields of GenAI such as “large language models” and “foundation models.” The Transformer architecture [3] and its pretraining bidirectional counterpart BERT [4] transformed the landscape of language representation to contextual embeddings that better model subtle semantic relationships. Transformer architecture [3] and its bidirectional pretraining variant BERT [4] changed the field towards contextual language representations that better model nuanced semantic relationships. Based on this, a hybrid neural topic model, BERTopic [5] is proposed, which preserves both the semantic coherence and the interpretability of topics by adopting both contextual embedding and class-based TF-IDF. However, the BERT-based methods are high computational cost, and the resulting topics are more difficult to validate and not easily able to be explicitly forecasted for trends. The parallel stream of work uses bibliometric knowledge graph analysis and keyword-co-occurrence analysis to create a GenAI research knowledge map at scale. While bibliometric reviews like those in [8] and [9] have been produced recently and describe the growing structure of GenAI research, they do so based on the intersection of multiple aspects of the research, namely, citations and keywords, and not on the actual meaning of the texts themselves which is then conveyed as a network of citations and keywords. Such bibliometric reviews [8] and [9] have been produced recently and describe the growing structure of GenAI research, but they do so by using the intersection of different aspects of the research; citations and keywords and not by using the meaning of the texts themselves, and they do not combine the statistical and contextual aspects of topic modeling into a single, interpretable pipeline. This gap is summarized in Table 1 – existing statistical and contextual approaches have limited ability, and bibliometric reviews lack semantic depth and predictive power.

Table 1. Comparison of existing knowledge-discovery approaches for Generative AI literature with the proposed hybrid framework

	Statistical Coherence	Contextual Understanding	Trend Prediction
LDA / LSA-based studies [1],[2]	Yes	No	No
BERT-based contextual models [3]-[5]	No	Yes	No
Bibliometric / keyword-co-occurrence reviews [8],[9]	Partial	No	Partial
This work (proposed hybrid)	Yes	Yes	Yes

Key Contribution

The novelty of this work is not just an algorithm, but a comprehensive and coherent pipeline for GenAI literature that integrates, in a comprehensible and forecasting way, the work of various NLP methods that were previously disjoint. The proposed framework includes a novel hybrid NLP framework that unites both statistical & contextual topic modeling; an improved topic coherence using cross-validating LDA/LSA topics and BERT semantic clusters; a context-aware knowledge-discovery process that mitigates lexical variation across GenAI subfields; and a precursor or base model to an emerging-topic prediction model that tracks topic evolution over time. In practice, the framework will enable automatic research analytics for institutions and funding agencies, an AI research-intelligence dashboard that shows the trend of thematics over time, decision support for research scholars who want to choose problem areas to work on that have growth potential and strategic technology forecasting for institutions who want to keep up with the GenAI landscape. These contributions complement the current knowledge base on literature-based knowledge discovery, in line with the goals of a contribution to the proceedings level of a systematic literature review.

Proposed Methodology

This is a proposal stage study so it is not an actual set of experimental results, but is a description of the methodology, which is planned for actual execution and validation, which will be reported in a future publication. The review protocol is followed by the guidelines for reporting systematic reviews by PRISMA 2020 [6] and the systematic literature review process suggested by Kitchenham and Charters [7] for planning, conducting and reporting the review.

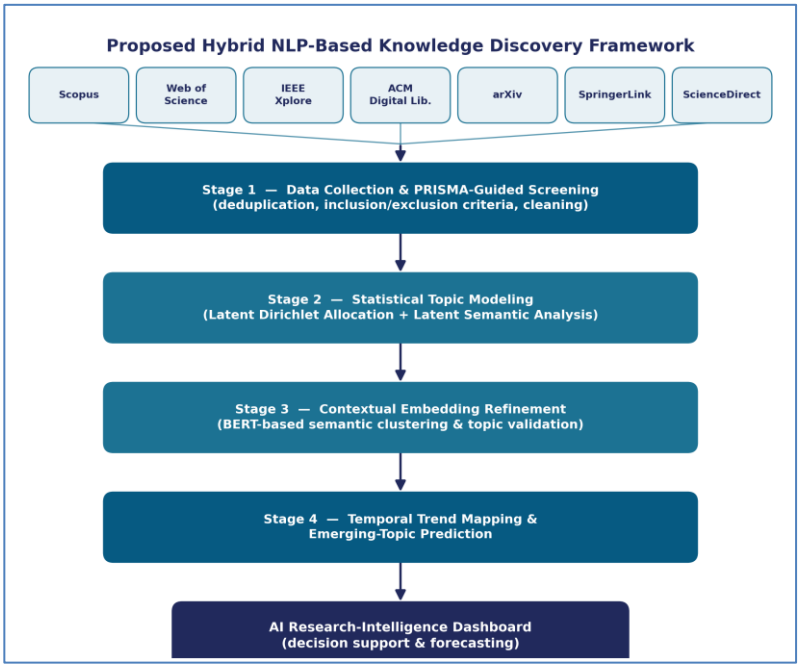

Figure 1. Proposed four-stage hybrid NLP-based knowledge discovery framework, integrating statistical topic modeling (LDA, LSA), contextual embedding refinement (BERT), and temporal trend mapping for emerging-topic prediction.

Seven databases will be used for complementary disciplines and coverage from the multidisciplinary and citation-quality perspective (Scopus, Web of Science), AI and engineering research (IEEE Xplore), computing and NLP-specific research (ACM Digital Library), emerging preprint research (arXiv), AI and data-science literatures (SpringerLink), and applied AI research (ScienceDirect). The article must be

written in English and published between 2018 and 2026, and must be peer reviewed, published in a journal, conference, or indexed preprint and pertain to Generative AI, Large Language Models or related topics with NLP. Exclusion criteria exclude non-peer-reviewed, duplicate, editorials and opinion pieces, and studies that do not focus on Generative AI. Search strings are made up of keywords like “Generative AI,” “large language models,” “topic modeling,” “knowledge discovery” and “natural language processing” linked by Boolean operators and formatted to suit the syntax of each database. According to the proposed hybrid framework (see Figure 1), there are four stages to this: First, the metadata (titles, abstracts, and keywords) are collected and then cleaned and pre-processed by deduplication, tokenization and removal of stop words. Second, LDA and LSA are used to obtain baseline topics from the cleaned corpus which are interpretable in terms of word co-occurrence statistics. Third, BERT-based context embeddings are computed for the same corpus and grouped to aid and semantically check the statistical topics, for instance when different lexica refer to the same concept. Fourth, topic assignments are monitored over publication years to visualize the development of topics over time, on which a new component of the topic prediction is built—flagging topics whose publication trends are increasingly upward. The general pipeline will be used to analyze the research-intelligence dashboard outlined in Section III.

Discussions

The literature reviewed in Section II shows a general trend that each individual technique is well developed and has been proven, but when used together, they are disjointed. The availability of large-scale literature, the development of transformer architectures, their increasing usage in industries, and the strong open-source community offer a promising context for automated knowledge discovery in the field of GenAI. Concurrently, the same literature also reports on existing deficiencies, such as incomplete discovery approaches, either statistical coherence or contextual understanding; but rarely both; lack of a common framework that incorporates several NLP paradigms; limited interpretability of deep-learning-based topic models; and poor performance when it comes to predicting new research trends. The goal of the design proposed in this paper is to directly overcome these limitations and not to provide a totally new modeling paradigm. The aim of the framework is to ensure that the topics models generated by LDA/LSA remain interpretable, as they are appealing to research administrators and reviewers, and to benefit from the contextual sensitivity of the embedding model BERT to correctly cluster semantically similar but lexically different GenAI terms. It is expected that, when tracking topic assignments over time, a static thematic map will become a forecasting tool, filling this gap that was uncovered in both topic modeling and bibliometric literature. As the framework is built upon already existing and tested components (LDA, LSA, BERT, PRISMA), there is not a significant amount of risk associated with the implementation of the framework, but the overall ability to implement it well will still rely on empirical assessment of topic coherence and forecasting accuracy with the set of completed literature in future research.

Conclusions

This paper is a proposal stage SLR that brings the following pointwise conclusions:

1. Problem statement addressed/motivation: The rapidly accelerating research on Generative AI has seen more than 18% annual increase in research activity [9] and researchers are unable to manually keep track of the themes emerging from this research, thus the need for an automated, interpretable and context-aware knowledge-discovery approach.
2. To close this gap, a systematic review protocol based on the PRISMA framework [6, 7] in seven databases, in which both a statistical topic modeling method (LDA, LSA) and a contextual embeddings method (BERT) were combined with each other in a hybrid approach, is suggested.
3. Key findings: A preliminary literature synthesis reveals that all of the mentioned statistical, contextual, and bibliometric approaches address only part of the knowledge-discovery problem

and none of them unites statistical coherence, contextual understanding and the ability to predict trends in one system, which is the gap that this study seeks to fill.

4. Limitations and future work: This paper proposes the framework and formulation of the problem, and corpus extraction, hybrid model implementation, and empirical evaluation of the validity of topic coherence and the forecasting accuracy are future tasks to be reported in a future publication.

References

1. D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003. https://jmlr.csail.mit.edu/papers/v3/blei03a.html
2. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas and R. A. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9
3. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, 2017, pp. 5998-6008. https://arxiv.org/abs/1706.03762
4. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter Assoc. Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, 2019, pp. 4171-4186. https://aclanthology.org/N19-1423/
5. M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv:2203.05794*, 2022. https://arxiv.org/abs/2203.05794
6. M. J. Page, J. E. McKenzie, P. M. Bossuyt, et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, n71, 2021. https://doi.org/10.1136/bmj.n71
7. B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," *EBSE Technical Report EBSE-2007-01*, Keele University and University of Durham, 2007.
8. S.-L. Ng and C.-C. Ho, "Generative AI in education: Mapping the research landscape through bibliometric analysis," *Information*, vol. 16, no. 8, art. 657, 2025. https://doi.org/10.3390/info16080657
9. H. Özyürek and U. Türen, "A bibliometric analysis of generative artificial intelligence in a multidisciplinary global perspective," *SN Computer Science*, vol. 7, no. 4, art. 318, 2026. https://doi.org/10.1007/s42979-026-04883-z