

A Comparative NLP Study of Semantic and Thematic Changes in AI Research Pre - and Post-generative AI

Inês Sena(✉), João Mendes, and Ana I. Pereira

Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
{ines.sena,joao.cmendes,apereira}@ipb.pt

Abstract. With the advancement of Artificial Intelligence (AI), particularly with the emergence of Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs), has significantly influenced the scope, accessibility, and dissemination of scientific research. This evolution raises important questions regarding how AI-related scientific discourse has changed over time. This study search lexical, semantic, and thematic changes in a dataset of approximately 50,000 AI-related publications retrieved from the Web of Science (WoS), comparing a pre-GenAI period (2015–2019) with a post-GenAI period (2022–2025). Natural Language Processing (NLP) techniques were applied, including TF-IDF for lexical analysis, Logistic Regression for temporal classification, and BERTopic for topic modeling. The results achieved an accuracy of 89%, revealing a clear separability between the two periods based on textual features. The analyses reveal a transition from predominantly technical and methodological terminology, focused on algorithms and infrastructure, toward more application-oriented and generative themes, including GenAI and LLMs. Furthermore, topic modeling shows the growing prominence of interdisciplinary domains such as education, healthcare, and business applications in the post-GenAI period. Overall, the findings suggest a significant lexical and thematic evolution in AI-related scientific discourse, reflecting broader transformations in research priorities and application domains following the emergence of GenAI technologies.

Keywords: Artificial Intelligence · GenAI · Natural Language Processing · Semantic Analysis · Topic Modeling

1 Introduction

Since 2022, a new Artificial Intelligence (AI) paradigm, namely Generative Artificial Intelligence (GenAI), experienced rapid growth with the emergence of tools such as ChatGPT and Gemini [6]. GenAI enables the generation of original content based on training over large-scale datasets containing text, images, audio, and video [10].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16759, pp. 272–288, 2027.
https://doi.org/10.1007/978-3-032-30518-3_18

GenAI is primarily driven by Large Language Models (LLMs), whose emergence has considerably transformed scientific communication and research dissemination. These tools now support tasks such as language editing, manuscript structuring, literature synthesis, text generation, and scientific reasoning. While offering new opportunities, they also raise important concerns regarding authorship, accountability, originality, and transparency in the scientific production process [5,8].

Current discussion extend beyond the technical capabilities of these systems. A growing body of literature points out that GenAI challenges traditional notions of scientific authorship, historically associated with human intellectual contribution, responsibility, and accountability. In this context, the relationship between human authors and GenAI should not be interpreted as replacement or opposition, but rather as an emerging form of techno-social collaboration that demands new conceptual and ethical considerations for academic production. [10].

At the same time, several studies show that the growing integration of GenAI into academic workflows provides operational benefits, including improved efficiency, support for non-native English speakers, acceleration of literature reviews, and assistance in research planning and analysis. However, these advantages coexist with persisting concerns regarding plagiarism, superficial analysis, bias, hallucinations, fabricated references, and loss of authorship. This reinforces the need for human supervision and transparent disclosure mechanisms regarding the use of these tools [5,8].

Furthermore, recent studies suggests that GenAI is no longer limited to supporting micro-level writing tasks but is increasingly influencing broader knowledge-production processes, including automated literature synthesis and support for scientific discovery. Nevertheless, the quality of generated outputs remains highly dependent on factors such as input data quality, prompt design, retrieval mechanisms, and human critical evaluation, which currently prevents the replacement of human scientific judgment [7,9].

In this context, it becomes relevant to explore not only how researchers use GenAI, but also how its emergence coincides with lexical, semantic and thematic changes in scientific discourse. This study addresses this challenge through the application of Natural Language Processing (NLP) techniques, which provide effective tools for analyzing large-scale textual corpora and identifying lexical patterns, semantic differences, and topic evolution over time.

This study searchs how scientific literature related to AI has evolved following the emergence of GenAI, through the analysis of a large-scale dataset of 50,000 publications extracted from the Web of Science (WoS) for two distinct periods: a pre-GenAI phase (2015–2019) and a post-GenAI phase (2022–2025). Several NLP techniques were applied, including TF-IDF for lexical analysis, supervised classification using Logistic Regression to distinguish periods, and BERTopic for topic modeling, aiming to identify structural lexical, semantic, and thematic changes in scientific discourse.

The main objective of this work is to study whether AI research has changed from predominantly technical topics toward more applied and gener-

ative approaches. To achieve this objective, the study adresses the following research questions (RQ):

- RQ1: How has the lexical and semantic representation of AI research evolved between the pre-GenAI (2015–2019) and post-GenAI (2022–2025) periods?
- RQ2: To what extent can NLP-based classification techniques h distinguish AI research between these two periods?
- RQ3: What tematic changes can be identified in AI research topics across the two periods?

The remainder of this paper is structured as follows. Section 2 reviews work on GenAI in scientific writing and on the use of NLP approaches to analyze the scientific literature. Section 3 describes the dataset and the data collection process. Section 4 presents the procedure, including pre-processing techniques and NLP approaches applied. Section 5 presents the results, showing the lexical, semantic, and thematic changes. Section 6 answers the research questions, comparing the results with the state-of-the-art findings. Finally, Sect. 7 concludes the article and describes future research directions.

The remainder of this paper is structured as follows. Section 2 reviews previous work on GenAI in academic contexts and the application of NLP techniques for scientific literature analysis. Section 3 describes the dataset and the data collection process. Section 4 presents the methodology, including preprocessing procedures and NLP techniques applied in the study. Section 5 presents the obtained results, showing lexical, semantic, and thematic changes across periods. Section 6 discusses the findings in relation to the research questions and the literature. Finally, Sect. 7 concludes the paper and outlines future research directions.

2 Related Work

This section presents the current context regarding the impact of Generative Artificial Intelligence (GenAI) on scientific communication and academic production, showing the main challenges, limitations, and research gaps identified in the literature. It also reviews the application of Natural Language Processing (NLP) approaches for the analysis of lexical, semantic, and thematic evolution in scientific discourse, providing the theoretical foundation for the methodological approach adopted in this study.

2.1 GenAI and Scientific Communication

Scientific communication practices have been increasingly influenced by the emergence of GenAI tools, as researchers progressively integrate these systems into academic workflows for tasks such as language improvement, literature review, text generation, research planning, and idea structuring [8]. Studies show benefits including improved efficiency, reduced writing effort, increased productivity, and better accessibility for non-native English speakers through translation and language assistance capabilities [8].

Due to the capabilities of Large Language Models (LLMs), GenAI tools are also being adopted in more advanced stages of scientific production, including automated literature synthesis, semantic exploration of large corpus, and support for knowledge discovery processes [9].

Despite these advantages, the growing use of GenAI raises significant ethical and methodological concerns regarding transparency, plagiarism, accountability, and research integrity. The use of AI-generated content challenges traditional notions of human intellectual contribution and scientific responsibility [10]. Since AI systems cannot assume accountability for scientific claims or conclusions, several publishers and academic institutions have explicitly rejected the attribution of authorship to AI systems [5].

Furthermore, the quality and reliability of AI-generated outputs remain strongly dependent on the quality of training data, prompt design, and human supervision [7]. Previous studies report that generated content may present superficial reasoning, fabricated references, hallucinations, limited originality, and embedded biases, reinforcing the importance of critical human evaluation and validation [5].

2.2 NLP for Scientific Literature Analysis

NLP has evolved considerably over recent decades, transitioning from rule-based and statistical approaches to Machine Learning (ML), Deep Learning, and transformer-based architectures [2]. Early techniques such as Word2Vec and GloVe enabled the representation of words through dense semantic vectors capable of capturing lexical similarity, although without contextual awareness [3]. More advanced transformer-based models, including BERT and related architectures, introduced contextual embeddings that substantially improved language understanding across multiple NLP tasks [3].

The emergence of Transformer architectures significantly advanced NLP capabilities by enabling models to capture long-range contextual dependencies and semantic relationships in large textual corpus [3]. These architectures showed improved performance in tasks such as text classification, semantic analysis, information extraction, document clustering, and topic modeling [3].

NLP techniques have increasingly been applied to the analysis of scientific literature, allowing researchers to extract structured information from large-scale unstructured corpus. Applications include information extraction, bibliometric analysis, trend identification, semantic similarity analysis, and topic modeling [1]. Embedding-based topic modeling approaches have also enabled the identification of temporal thematic evolution across scientific domains [4].

Recent studies further reveal the growing role of NLP in supporting automated knowledge discovery from scientific literature, particularly in tasks involving literature review automation, semantic exploration, and trend detection [4]. These capabilities make NLP especially suitable for investigating how scientific discourse evolves in response to technological transformations such as the emergence of GenAI.

2.3 Research Gap

Although studies extensively discuss the applications, opportunities, and limitations of GenAI in academic contexts, comparatively less attention has been given to understanding how the emergence of GenAI is reflected in the evolution of scientific discourse itself. Most previous research focuses either on ethical implications or on the direct use of AI-assisted writing tools, rather than on broader lexical, semantic, and thematic transformations in scientific literature.

Additionally, current literature lacks large-scale empirical analyses combining NLP techniques with longitudinal scientific corpus to search how AI-related research discourse has evolved before and after the emergence of GenAI technologies. In particular, there remains limited evidence regarding whether recent changes observed in AI literature reflect a broader transition from technically oriented research toward more applied, interdisciplinary, and generative-oriented topics.To address this gap, the present study applies multiple NLP techniques to a large corpus of scientific publications extracted from the Web of Science (WoS), comparing pre-GenAI and post-GenAI periods to identify lexical, semantic, and thematic changes in AI-related scientific discourse.

3 Data Characterization

This study is based on a large-scale dataset of scientific publications obtained from the WoS database. The dataset focuses on research related to AI and was divided into two distinct temporal periods to enable a comparative analysis between pre- and post-GenAI research trends. Figure 1 presents the data collection and processing workflow adopted in this study.

The first period, referred to as the pre-GenAI phase, covers the years between 2015 and 2019 and represents the state of AI-related scientific research prior to the widespread adoption of GenAI technologies. The second period, referred to as the post-GenAI phase, covers the years between 2022 and 2025, reflecting the rapid expansion of GenAI and LLMs in scientific and technological contexts.

The years 2020 and 2021 were intentionally excluded from the analysis to create a clearer separation between the two phases and reduce overlap between emerging and consolidated GenAI-related discourse. These years correspond to a transitional period characterized by the increasing popularity of Transformer-based architectures and the emergence of early LLMs, such as chatGPT-3, which had not yet achieved broad integration into scientific communication practices.

For each period, 25,000 publications were extracted from WoS. Due to the platform limitation of exporting a maximum of 1,000 records per extraction, the data collection process was performed in multiple stages, generating several files that were subsequently merged into a unified dataset containing approximately 50,000 publications.

The WoS database provides metadata fields including title, abstract, keywords, authors, year of publication, source title, and document identifiers (e.g., DOI) for each publication. However, for the Natural Language Processing

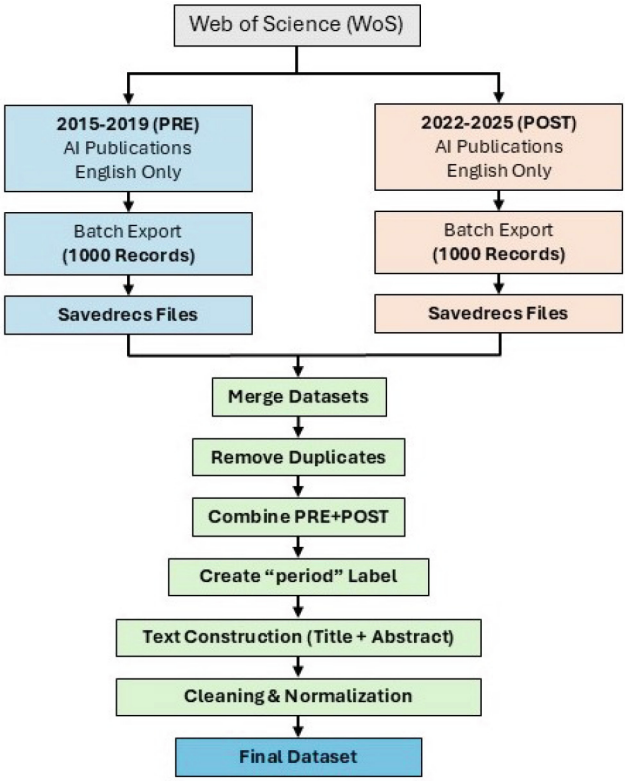

Fig. 1. Data collection and processing workflow.

(NLP) analysis conducted in this study, only textual fields, primarily titles and abstracts, were considered.

Following data collection, the dataset underwent a preprocessing stage involving duplicate removal, handling of missing values, and textual normalization. Additionally, a categorical variable indicating the corresponding temporal period (pre-GenAI or post-GenAI) was assigned to each document to support comparative analysis.

The final dataset contains structured metadata fields, including publication year, title, abstract, keywords, DOI, document type, source title, and source file information, as well as a processed textual field combining titles and abstracts. After preprocessing, the dataset contained 24,781 documents for the pre-GenAI period and 24,865 documents for the post-GenAI period.

The resulting corpus was subsequently prepared for NLP analysis and served as the basis for the lexical analysis, supervised classification, and topic modeling procedures described in the following sections.

4 Methodology

This study uses a multi-stage Natural Language Processing (NLP) pipeline to analyze lexical, semantic, and thematic differences between AI-related scientific publications from the pre-GenAI and post-GenAI periods. The methodology combines traditional statistical text analysis techniques with recent embedding-based approaches for document classification and topic modeling. Figure 2 presents the methodological workflow adopted in this study.

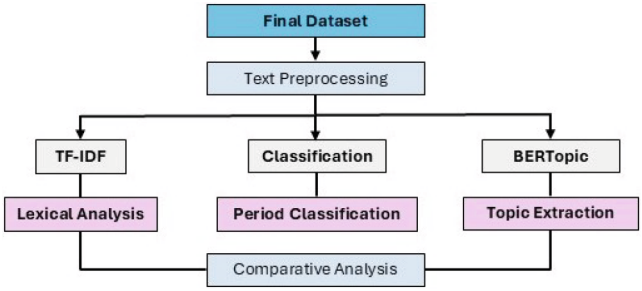

Fig. 2. Methodological pipeline used.

4.1 Text Preprocessing

The textual corpus was constructed by combining the title and abstract of each publication into a single textual field. Before analysis, several preprocessing procedures were applied to improve data consistency and reduce textual noise. These procedures included lowercasing, removal of punctuation and special characters, normalization of whitespace, and elimination of incomplete or duplicated records.

The preprocessing stage is particularly important in NLP tasks, as it reduces lexical variability and improves the quality of feature extraction and semantic representations [2].

4.2 Lexical Analysis Using TF-IDF

To identify the most representative lexical patterns within each temporal period, the Term Frequency–Inverse Document Frequency (TF-IDF) method was applied. TF-IDF is a statistical weighting approach that evaluates the relevance of terms based on their frequency within a document while reducing the importance of highly frequent words across the entire corpus.

TF-IDF is widely used in text mining and information retrieval due to its interpretability and effectiveness in identifying discriminative lexical features [2]. In this study, TF-IDF was employed to compare dominant terms between the pre-GenAI and post-GenAI corpus, enabling the identification of lexical modifications associated with changes in AI-related scientific discourse.

4.3 Supervised Classification

A supervised classification approach was used to evaluate whether scientific publications from the two temporal periods could be automatically distinguished based on their textual characteristics. Each document was labeled according to its corresponding period (pre-GenAI or post-GenAI), and textual features were extracted using TF-IDF vectorization.

The dataset was divided into training and testing subsets using a 70/30 split strategy, where 70% of the documents were used for training and 30% for evaluation. To reduce the risk of overfitting and improve generalization, the Logistic Regression classifier was selected due to its robustness, interpretability, and suitability for sparse and high-dimensional textual data [3]. In addition, regularization mechanisms inherent to Logistic Regression were maintained during training to penalize excessively complex decision boundaries and improve model stability.

The classification model was evaluated using unseen test data through metrics including accuracy, precision, recall, F1-score, and confusion matrix analysis. The confusion matrix was particularly important for identifying classification balance and potential misclassification patterns between the two periods.

Beyond predictive performance, the model coefficients were also analyzed to identify the most discriminative terms associated with each period. This procedure provided additional insights into lexical and semantic differences between pre-GenAI and post-GenAI AI-related scientific publications.

4.4 Topic Modeling with BERTopic

To analyse thematic evolution across the two periods, BERTopic was applied as the primary topic modeling technique. BERTopic combines Transformer-based embeddings with clustering algorithms to generate semantically coherent topics from large textual corpora.

Document embeddings were generated using a pre-trained SentenceTransformer model, which produces dense contextual vector representations capable of capturing semantic relationships beyond isolated lexical features [3]. Compared to traditional bag-of-words representations, contextual embeddings provide improved semantic understanding of complex scientific text.

BERTopic leverages these embeddings together with dimensionality reduction and density-based clustering techniques to identify groups of semantically related documents. This approach is particularly suitable for dynamic and multidisciplinary domains such as AI research, where terminology and thematic structures evolve rapidly over time [4].

The identified topics were subsequently analyzed and compared between periods to observe thematic transitions associated with the emergence and expansion of GenAI-related research.

4.5 Comparative Analysis

To analyze changes in terms, thematic focus, and research direction associated with the emergence of GenAI, results from the two periods were compared with respect to the extracted variables, classification outcomes, and topic distributions.

The combined methodological approach enables the integration of interpretable lexical analysis with more advanced embedding-based semantic representations, providing a comprehensive framework for studying the evolution of AI-related scientific discourse before and after the emergence of GenAI technologies.

5 Results

This section presents the results obtained from applying the proposed NLP pipeline to the AI-related scientific corpus. The findings are organized into three main analytical components: lexical analysis using TF-IDF, supervised classification, and topic modeling using BERTopic.

5.1 Lexical Analysis Using TF-IDF

To analyze lexical differences between the pre-GenAI (2015–2019) and post-GenAI (2022–2025) periods, TF-IDF was applied to identify the most representative terms in each corpus. Table 1 presents the top 15 terms with the highest average TF-IDF values for each period.

The results reveal clear lexical differences between the two periods. In the pre-GenAI period, the most representative terms are strongly associated with technical and methodological aspects of AI research, including neural networks, algorithms, machine learning, and model development. Terms such as *learning*, *neural*, *network*, and *algorithm* indicate a stronger focus on foundational AI methods and computational infrastructure.

In contrast, the post-GenAI period presents a broader and more application-oriented vocabulary. Terms such as *applications*, *analysis*, *review*, and *research* become more prominent, suggesting an expansion of AI research toward interdisciplinary and applied domains. In addition, GenAI-related terminology, including references to LLMs and generative systems, emerged more frequently during this period.

Overall, these findings suggest a lexical transition from technically centered AI research toward a more diversified scientific discourse characterized by broader application contexts and interdisciplinary integration.

5.2 Supervised Classification

To evaluate the separability between the two temporal periods, a supervised classification model based on Logistic Regression and TF-IDF features was trained

Table 1. The top 15 of the most representative terms in the pre-GenAI (2015–2019) and post-GenAI (2022–2025) periods.

Term PRE (2015–2019)	Mean TF-IDF	Term POST (2022–2025)	Mean TF-IDF
learning	0.032	ai	0.060
artificial	0.028	learning	0.024
data	0.027	data	0.023
intelligence	0.027	research	0.022
artificial intelligence	0.024	study	0.020
ai	0.022	review	0.017
model	0.0211	systems	0.017
based	0.0198	models	0.017
neural	0.0197	use	0.016
using	0.0196	intelligence ai	0.016
network	0.0194	analysis	0.016
algorithm	0.0194	applications	0.016
machine	0.0183	using	0.016
paper	0.0176	development	0.015
systems	0.0173	model	0.015

and tested. The dataset was divided using a 70/30 split, where 70% of the documents were used for training and 30% for testing. Table 2 presents the classification results obtained on the test dataset.

Table 2. Classification results for separability between the two periods.

Label	Accuracy	Recall	F1-score	Support
2015–2019 (0)	0.882	0.894	0.888	7392
2022–2025 (1)	0.894	0.881	0.887	7460

The results in Table 2 show an accuracy of 89%, revealing a strong distinction in the terms used between pre- and post-GenAI publications. Furthermore, they show a balance between classes, revealing consistent classification performance, as the accuracy and recall values presented are balanced between the two periods. To provide a more detailed analysis of classification performance. Figure 3 presents the confusion matrix obtained from the test dataset.

The confusion matrix shows that the majority of documents were correctly classified in their corresponding periods, with relatively low misclassification rates between classes. Specifically, most pre-GenAI documents were correctly

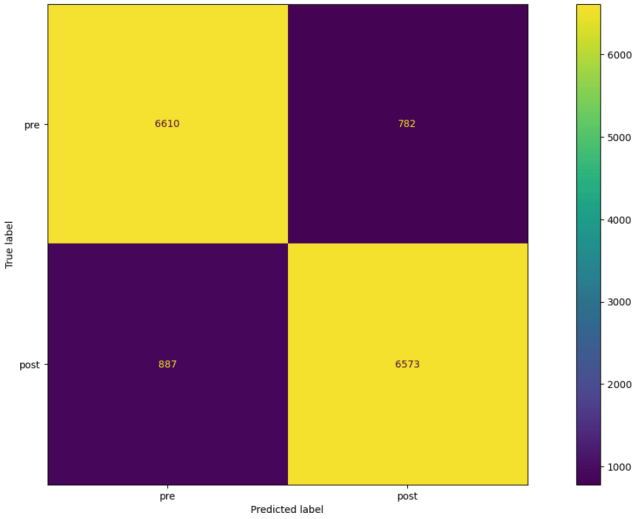

Fig. 3. Confusion matrix of Logistic Regression results to classify terms in pre-GenAI (2015–2019) and post-GenAI (2022–2025).

identified as belonging to the 2015–2019 period, while most post-GenAI publications were correctly assigned to the 2022–2025 period. The balanced distribution of false positives and false negatives suggests that the lexical differences between periods are sufficiently robust to support reliable automatic classification.

A more detailed examination of the classifier coefficients further revealed distinct discriminative vocabularies associated with each period. Publications from the pre-GenAI phase were more strongly associated with terms related to traditional AI methodologies, ML infrastructure, and algorithmic development. In contrast, publications from the post-GenAI phase contained a greater prevalence of terms associated with GenAI, emerging technologies, interdisciplinary applications, and broader societal integration of AI systems.

These findings reinforce the hypothesis that AI-related scientific discourse underwent substantial lexical and semantic changes following the emergence of GenAI technologies.

5.3 Topic Modeling with BERTopic

To analyze thematic evolution across periods, BERTopic was applied to the corpus to identify latent semantic topics based on documents embedding and clustering.

Figure 4 presents the most significant thematic changes between the pre-GenAI (2015–2019) and post-GenAI (2022–2025) periods.

The BERTopic analysis reveals a substantial increase in application-oriented and interdisciplinary topics during the post-GenAI period. Educational themes, including *education*, *students*, and *teaching*, showed some of the largest increases,

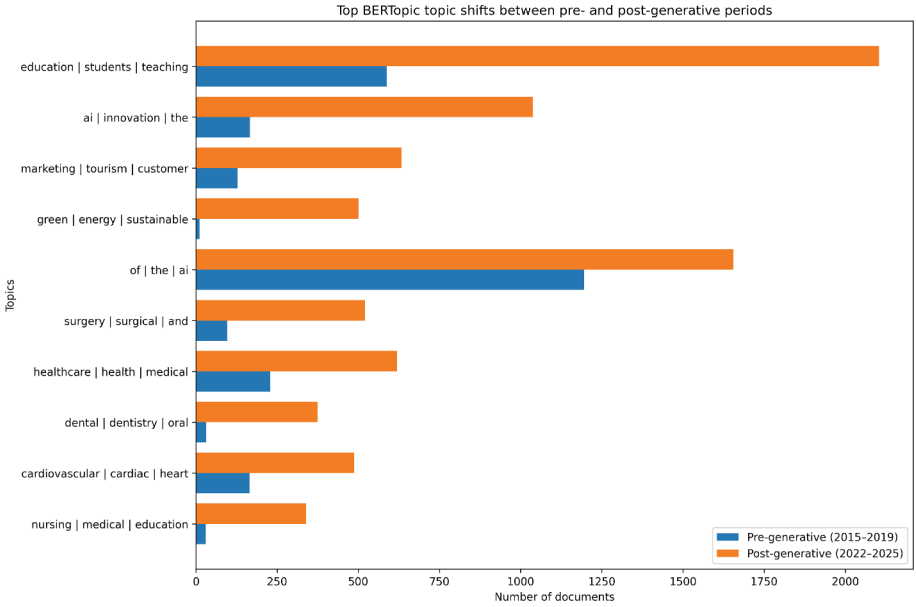

Fig. 4. Comparison of the most significant topic changes between pre-GenAI (2015–2019) and post-GenAI (2022–2025) periods.

suggesting a growing integration of AI technologies within educational environments.

Similarly, topics associated with business, innovation, and marketing expanded considerably, indicating a broader industrial and commercial adoption of AI systems. Medical and healthcare-related topics, including *health*, *surgery*, *dentistry*, and *cardiovascular* applications, also showed strong growth, reflecting the increasing role of AI in clinical and healthcare research.

Environmental and sustainability-related themes, such as *green energy* and *sustainability*, became more prominent in the post-GenAI period, further emphasizing the diversification of AI applications into new scientific and societal domains.

Although foundational AI topics remained present across both periods, their relative prominence was lower compared to the rapid growth of application-oriented themes. Overall, the BERTopic results suggest a thematic transition from predominantly technical and methodological research toward a broader interdisciplinary and application-focused research landscape.

These findings are consistent with the lexical and classification analyses, collectively supporting the substantial structural changes in AI-related scientific discourse following the emergence and expansion of GenAI technologies.

6 Discussion

The results of this study provide consistent evidence of a transformation in AI-related scientific discourse, particularly following the emergence and growth of GenAI. By combining lexical analysis, supervised classification, and topic modeling, it was possible to identify both surface-level linguistic changes and broader structural transformations in research themes and scientific communication patterns across the two analyzed periods.

The outcomes suggest that the evolution of AI technologies is reflected not only in the topics being study but also in the vocabulary, semantic structures, and research orientation of scientific publications. In addition, the results reinforce the idea that the rise of GenAI coincides with a transition from predominantly methodological and infrastructure-oriented research toward broader interdisciplinary and application-driven scientific production.

6.1 RQ1 How Has the Lexical and Semantic Representation of AI Research Evolved Between the Pre-GenAI (2015–2019) and Post-GenAI (2022–2025) Periods?

The TF-IDF analysis revealed substantial lexical differences between the two periods. In the pre-GenAI phase, the dominant vocabulary was strongly associated with technical and methodological aspects of AI research, including terms related to algorithms, neural networks, ML models, and computational systems. This pattern reflects a research landscape primarily focused on the development and optimization of AI techniques and infrastructure.

In contrast, the post-GenAI period presented a broader and more application-oriented vocabulary. Terms associated with healthcare, education, management, analysis, and applications became considerably more prominent, suggesting an expansion of AI research into interdisciplinary and real-world domains. Furthermore, the emergence of terms directly associated with GenAI and LLMs, including references to generative systems and conversational AI tools such as ChatGPT, show the growing influence of these technologies on scientific discourse.

These outcomes are consistent with previous studies describing the expansion of NLP and LLM applications across multiple scientific domains [3,4]. However, while most previous literature focuses on technological capabilities and ethical implications, this study contributes empirical evidence showing how these technological developments are reflected in the language used in scientific publications.

Nevertheless, the results also reveal an important distinction between thematic evolution and stylistic transformation. Although the study identifies significant lexical and semantic changes, these changes are mainly associated with changes in research topics and application domains rather than with deeper stylistic characteristics of scientific writing itself. Since the analysis was limited to titles and abstracts, additional stylistic modifications present in full-text sec-

tions such as methodology, discussion, and conclusions may not have been fully captured.

6.2 RQ2: To What Extent Can NLP-Based Classification Techniques H Distinguish AI Research Between These Two Periods?

The supervised classification results demonstrate that publications from the two periods can be reliably distinguished based solely on textual features. The Logistic Regression classifier achieved approximately 89% accuracy, showing that the lexical and semantic differences between periods are sufficiently strong to support automatic temporal classification.

The confusion matrix further confirmed the robustness of the model, showing balanced classification performance across both classes and relatively low misclassification rates. The similarity between precision, recall, and F1-score values suggests that the classifier generalized consistently and that overfitting effects were minimized through the use of TF-IDF vectorization, regularized Logistic Regression, and the separation of independent training and testing datasets.

These results reinforce previous research demonstrating the effectiveness of NLP techniques in identifying patterns and trends within large scientific corpus [2]. More importantly, the outcomes reveal that the growth of GenAI corresponds to measurable linguistic and semantic transformations in scientific communication.

Therefore, while the classification results clearly demonstrate the existence of temporal linguistic differences, they do not independently establish causality between GenAI-assisted writing and the observed discourse transformations.

6.3 RQ3: What Tematic Changes Can Be Identified in AI Research Topics Across the Two Periods?

The BERTopic analysis provided strong evidence of thematic evolution between the pre-GenAI and post-GenAI periods. While the earlier period was dominated by topics associated with technical AI development and computational methodologies, the later period showed a substantial diversification of research themes.

The most significant growth was observed in application-oriented domains such as education, healthcare, business, sustainability, and interdisciplinary AI adoption. These outcomes indicate that AI research increasingly extends beyond core methodological development toward broader societal, industrial, and scientific applications.

This thematic diversification is aligned with recent literature showing that transformer-based NLP methods and embedding-based topic modeling approaches are capable of identifying long-term research evolution patterns within scientific corpusa [1,4]. The present study extends these results by specifically showing how the emergence of GenAI technologies coincides with a measurable expansion of interdisciplinary AI-related research topics.

At the same time, the results suggest that AI is progressively transitioning from a specialized technical discipline into a general-purpose scientific and technological paradigm integrated across multiple domains.

6.4 Implications and Research Contributions

In general, the results suggest a paradigmatic change in AI research, characterized by a transition from method-centered development to generative and application-oriented approaches. This change is not reflected only in the topics studied but also in the language used by researchers.

From a methodological perspective, this study shows the effectiveness of combining classical NLP approaches with modern embedding-based approaches for the analysis of large scientific corpora. Furthermore, it enhances the growing corpus of literature on the impact of GenAI on scientific writing by supplying quantitative evidence of changes in discourse patterns, rather than focusing only on conceptual or moral discussions.

Overall, the findings suggest the existence of a paradigmatic transformation in AI-related scientific discourse following the emergence of GenAI technologies. The results indicate a transition from predominantly technical and method-centered research toward broader application-oriented and interdisciplinary approaches.

From a methodological perspective, this study demonstrates the effectiveness of combining classical NLP techniques, such as TF-IDF and supervised classification, with modern embedding-based topic modeling approaches for analyzing large-scale scientific corpus. The integration of these methods enabled the identification of lexical, semantic, and thematic changes across distinct temporal periods.

The study also contributes to the growing literature on GenAI and scientific writing by providing quantitative evidence of changes in scientific discourse patterns. While previous research has primarily focused on ethical concerns, authorship issues, and conceptual discussions surrounding GenAI, this work offers empirical evidence of measurable transformations in AI-related scientific communication.

7 Conclusion and Future Works

This study explore the evolution of AI-related scientific research before and after the emergence of GenAI. Using a large-scale dataset of approximately 50,000 publications extracted from the Web of Science (WoS), a combination of NLP techniques, including TF-IDF, supervised classification, and BERTopic, was applied to identify lexical, semantic, and thematic changes in scientific discourse across two distinct periods.

The results provide consistent evidence of a significant transformation in AI research following the emergence of GenAI technologies. The lexical analysis

revealed a transition from predominantly technical and methodological terminology toward a broader and more application-oriented vocabulary. Terms associated with infrastructure, algorithms, neural networks, and model development became progressively accompanied by concepts related to education, healthcare, business, sustainability, and generative systems.

The supervised classification results further showed that the two periods can be reliably distinguished based solely on textual characteristics, confirming the existence of systematic linguistic differences between pre- and post-GenAI scientific publications. In addition, the BERTopic analysis revealed a thematic diversification of AI research, showing a change from method-centered investigations toward interdisciplinary and application-driven domains strongly associated with the use of GenAI and LLMs.

Taken together, these results suggest the paradigmatic transformation in AI-related scientific discourse. The outcomes indicate that the rapid development and widespread use of GenAI technologies are influencing not only the topics being searched but also the way scientific knowledge is communicated and structured within the research community.

From a methodological perspective, this study shows the effectiveness of combining classical NLP techniques with modern embedding-based approaches for the large-scale analysis of scientific literature. The proposed framework provides a replicable methodology for studying temporal changes in research domains and scientific communication patterns.

Despite these contributions, several limitations should be acknowledged. First, the analysis was limited to titles and abstracts, which may not fully capture deeper methodological, theoretical, or stylistic transformations present in complete scientific papers. Second, the exclusion of the 2020–2021 transition period intentionally created a clearer separation between the pre- and post-GenAI phases but may have omitted important intermediate developments associated with the early use of transformer-based models such as chatGPT-3. Third, although the results reveal strong correlations between the emergence of GenAI and changes in scientific discourse, they do not establish direct causal relationships between AI-assisted writing tools and the observed linguistic transformations.

Future research may address these limitations by incorporating full-text analysis, stylistic and syntactic metrics, citation networks, and longitudinal analyses including transitional periods. Additional studies may also integrate other scientific databases and explore more advanced semantic approaches based on LLMs to obtain deeper contextual representations of scientific discourse. Furthermore, future investigations could focus more directly on the influence of GenAI tools on authorship practices, writing styles, scientific originality, and transparency in academic communication.

Overall, this study contributes to a broader understanding of how emerging AI technologies are reshaping scientific research and communication, providing a foundation for future interdisciplinary investigations into the long-term impact of GenAI on scientific production.

Acknowledgments. This work was supported by FCT- Fundação para a Ciência e Tecnologia, I.P. by projects: CeDRI, UID/05757/2025 (DOI: 10.54499/UID/05757/2025) and UID/PRR/05757/2025 (DOI:10.54499/UID/PRR/05757/2025); LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020). This work has been supported by project Ted-SOEP - 101179402 and project Di-Hor - 2025-1-IT02-KA220-SCH-000359420.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article. During the preparation of this work, the authors used ChatGPT-5 to improve the paper's writing and clarity. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

1. Natural language processing in construction: A scientometric review. Automation in Construction (2022). https://doi.org/10.1016/j.autcon.2022.104169
2. Natural language processing for low-resource languages: A survey. Natural Language Processing (2024). https://doi.org/10.1017/nlp.2024.33
3. A comprehensive review of bert and its applications in nlp. Artificial Intelligence Review (2025). https://doi.org/10.1007/s10462-025-11162-5
4. Natural language processing for materials science. npj Computational Materials (2025). https://doi.org/10.1038/s41524-025-01554-0
5. Chauhan, C.: The impact of generative artificial intelligence in scientific content synthesis for authors. Am. J. Pathol. **194**(8), 1406–1408 (2024)
6. Dhamani, N., Engler, M.: Introduction to generative AI. Simon and Schuster (2026)
7. Leong, S.X., Loh, N.D.: Writing scientific review articles in the age of generative artificial intelligence (2025)
8. Meishar-Tal, H.: Scientific writing in the age of generative ai: usage patterns and ethical considerations among israeli researchers. Ethics & Behav., 1–16 (2025)
9. Reddy, C.K., Shojaee, P.: Towards scientific discovery with generative ai: progress, opportunities, and challenges. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 28601–28609 (2025)
10. Watson, S., Brezovec, E., Romic, J.: The role of generative ai in academic and scientific authorship: an autopoietic perspective. AI & Soc. **40**(5), 3225–3235 (2025)