



GraphiBERT-ML: A Knowledge-Enhanced NER Approach for Cross-Domain Comparative Analysis of Machine Learning Literature

Nabila Khouya^{1*}, Asmaâ Retbi², Samir Bennani³

^{1,2,3}Rime Team, Mohammadia School of Engineers (EMI), Mohammed V University in Rabat, Rabat, Morocco

*nabi97ky@gmail.com

Abstract

The exponential growth of scientific literature on platforms such as arXiv presents a major challenge in identifying and comparing key contributions to machine learning across diverse academic domains. To address this, we propose GraphiBERT-ML, a knowledge-enhanced extension of BERT that integrates semantic embeddings extracted from DBpedia to improve named entity recognition (NER) in scientific articles. To the best of our knowledge, this study presents the first knowledge-enhanced NER model that explicitly integrates DBpedia-based embeddings for large-scale cross-domain scientific analyses. The model was evaluated on a cross-domain dataset spanning eight fields, including computer science, physics, biology, finance, and economics. Experimental results show that GraphiBERT-ML achieves its highest performance in computer science, with an accuracy of 0.9372, an F1-score of 0.9368, and a precision of 0.9376. Physics and mathematics also demonstrate strong performance (F1-scores of 0.9115 and 0.8970), while more heterogeneous domains such as biology and finance show lower scores (F1-scores of 0.7946 and 0.7872), reflecting the complexity and variability of their terminology. Across all domains, GraphiBERT-ML consistently outperformed the baseline BERT model, confirming the benefit of external knowledge integration for scientific NER. These findings highlight domain-specific challenges in entity extraction and demonstrate the potential of knowledge-augmented models to advance cross-disciplinary analysis of machine learning research.

Keywords: academic research; language models; machine learning; natural language processing; named entity recognition

How to Cite: N. Khouya, A. Retbi, and S. Bennani, “GraphiBERT-ML: A Knowledge-Enhanced NER Approach for Cross-Domain Comparative Analysis of Machine Learning Literature”, *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 2, pp. 280 - 288, Apr. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i2.7160>

Received: September 7, 2025

Accepted: March 3, 2026

Available Online: April 6, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

In an academic world where knowledge is multiplying exponentially [1], [2], the capacity to extract and integrate information has become essential. The arXiv platform [3], with its vast collection of articles spanning computer science, finance, biology, and many other disciplines, represents an invaluable resource for scientific exploration. However, given the overwhelming volume of available information [4], [5], effective methods are needed to extract meaningful insights, particularly within the field of machine learning [6].

Machine learning [7], and especially named entity recognition (NER) tasks [8], plays a key role in analyzing and processing large-scale textual data. Yet, systematically comparing machine learning contributions across disciplinary domains such as physics, mathematics, finance, or biology remains a major challenge [9], [10]. Such cross-domain comparison is crucial for identifying similarities and

differences in methodological choices, application contexts, and research priorities across diverse fields [11].

To address this challenge, this article proposes an innovative approach for extracting, annotating [12], and comparing arXiv articles that intersect machine learning with various disciplinary domains. We develop a novel annotation strategy combined with a dual-filtering mechanism, and introduce GraphiBERT-ML, an enhanced version of BERT that incorporates external semantic knowledge extracted from DBpedia. By filtering and analyzing articles based on rigorous criteria—such as machine learning task types, software tools, and reported performance—we aim to map the current landscape of machine learning within these scientific fields [13].

Several recent studies have explored the application of advanced natural language processing techniques for named entity recognition in scientific corpora. A relevant line of research is the Fast-NER strategy [14], which applies large language models (LLMs) such as

Claude-2 [15], LLaMA-2-70B [16], and GPT-3.5 [17] to astronomy documents. This approach demonstrates that LLMs can extract named entities through prompt engineering without domain-specific training. However, while effective within a single field, Fast-NER does not consider cross-domain variability nor incorporate external structured knowledge—two key challenges when analyzing machine learning literature spanning multiple scientific disciplines.

Similarly, approaches related to Prompt-NER apply advanced NLP techniques to annotate scientific text using prompt-based strategies. While these methods show promising results in domain-specific contexts, they remain largely confined to individual disciplines and do not investigate how machine learning entities vary across heterogeneous domains such as computer science, biology, finance, or physics. In addition, these approaches do not explore the integration of external structured knowledge sources to enhance NER accuracy.

Another line of research combines NLP with knowledge graphs in healthcare applications, where clinical narratives are mapped to ontological concepts [18]. This work demonstrates the potential of structured semantic knowledge for improving text understanding. However, the approach is limited to biomedical data and focuses primarily on ontology completion rather than enhancing NER performance in large-scale interdisciplinary scientific corpora.

Taken together, these studies demonstrate important advances in NER through prompt-based methods, large language models, and knowledge graph integration. Nevertheless, significant challenges remain when considering large-scale interdisciplinary scientific corpora. Fast-NER and Prompt-NER approaches are primarily evaluated within single-domain settings and do not systematically assess cross-domain variability. Similarly, knowledge graph integration methods are typically domain-specific and focus mainly on ontology alignment rather than improving NER performance across heterogeneous academic disciplines.

Furthermore, most transformer-based NER frameworks rely exclusively on contextual textual embeddings and rarely incorporate structured semantic knowledge directly into their architectures. Although knowledge graphs have been explored in specialized domains, their integration into transformer-based NER models for large-scale cross-domain scientific analysis remains underexplored. In particular, the impact of structured knowledge embeddings on interdisciplinary machine learning literature has received limited investigation.

In this context, DBpedia provides a structured, domain-spanning knowledge base capable of enriching contextual token representations with semantic information applicable across multiple scientific fields. For large-scale interdisciplinary corpora such as arXiv, DBpedia offers broad domain coverage and a well-

structured ontology, making it particularly suitable for cross-domain scientific analysis compared to knowledge bases that are restricted to specific application areas. By integrating DBpedia-based knowledge embeddings into a transformer architecture, GraphiBERT-ML offers a unified framework for domain-aware and cross-disciplinary NER in scientific machine learning corpora.

Despite substantial progress in transformer-based NER models and prompt-driven large language models, several limitations remain. Existing approaches are often confined to single-domain corpora, which limits their ability to generalize across heterogeneous scientific disciplines. Furthermore, previous research has not yet systematically addressed the integration of structured semantic knowledge within transformer-based architectures for large-scale cross-domain analysis of machine learning literature. As a result, the comparison of machine learning entities across diverse academic fields remains methodologically constrained.

Building on recent advances in large language models such as BERT, RoBERTa, and DistilBERT, our work focuses on named entity recognition in machine learning literature. In our previous study, BERT demonstrated superior performance for NER in scientific articles compared with other transformer-based models. In this paper, we extend this line of research by introducing GraphiBERT-ML, a knowledge-enhanced variant of BERT that integrates structured semantic information from DBpedia. To the best of our knowledge, this work is among the first to systematically integrate DBpedia-based knowledge embeddings into a transformer architecture for large-scale cross-domain scientific NER.

Rather than relying solely on raw textual representations or prompt-based LLM strategies, GraphiBERT-ML combines contextual token embeddings with structured external knowledge, enabling more robust and domain-aware analysis of machine learning entities across disciplines such as computer science, biology, finance, and physics.

This study is guided by two main research questions. The first examines how the systematic analysis of machine learning articles—particularly through NER applied across multiple disciplinary domains—can deepen our understanding of emerging trends, patterns, and innovations within these fields. The second investigates the extent to which this analytical framework enables an effective comparison of NER-related contributions across domains as diverse as computer science, physics, mathematics, biology, and finance, as well as other areas represented on arXiv.

This work makes several contributions. First, we introduce a rigorous method for identifying and annotating arXiv articles that address both machine learning and multiple disciplinary domains [19]. Second, we present GraphiBERT-ML, which

significantly improves entity recognition performance by integrating pre-trained language models with external domain knowledge. This approach enables a comparative analysis of contributions, highlighting similarities and differences in methodologies and applications across these scientific fields [20]. Our results indicate that computer science exhibits the highest performance metrics, reflecting the importance of structured data and mature ML practices. Through systematic analysis and extensive annotation, we address the difficulties of cross-domain comparison and provide new perspectives on the interconnected nature of scientific knowledge, opening opportunities for interdisciplinary collaboration and large-scale automation in scientific literature analysis.

2. Methods

In order to achieve an enriching cross-domain comparison of contributions to machine learning, our study relies on a detailed annotation process of articles

from arXiv. This approach focuses on the bidimensional filtering strategy, where each article is selected according to its affiliation to a main domain such as physics and its participation in machine learning. This dual filter approach ensures that our analysis focuses solely on work that intersects machine learning with specific scientific fields, as shown in Figure 1, thus enabling an objective and contextual assessment of technological and methodological advances.

Figure 1 illustrates a structured process for the cross-domain comparison of contributions to machine learning. It begins with an initial filtering of articles based on their primary domain, such as physics, biology, etc. Each domain then undergoes an annotation stage, followed by the application of specific algorithms. The results from each domain are then compared cross-domain to provide an enriched and contextualized analysis of technological and methodological progress in machine learning research.

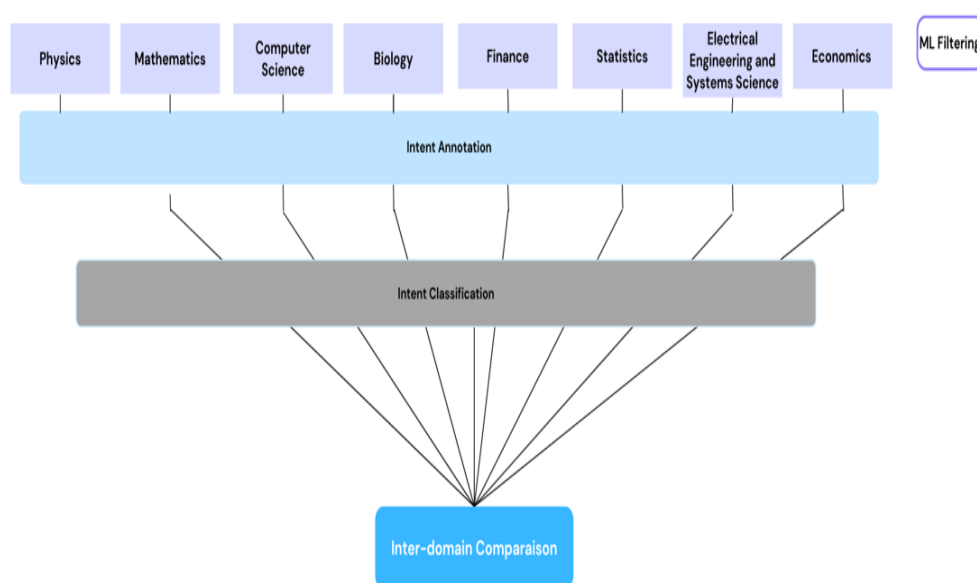


Figure 1. Cross-Domain Comparison in Machine Learning

To illustrate the entire process, we have also developed a diagram of the proposed architecture as shown in Figure 2. Each step of this architecture will be described thoroughly in subsequent sections.

To offer a clear summary of the dataset, Table 1 presents the distribution of articles collected across eight scientific domains. For each domain, it reports the number of articles, the total number of segmented sentences, and the average sentence count per article. This last indicator reflects the typical length of publications within each field. Taken together, these statistics provide essential context for understanding the methodological workflow illustrated in Figures 1 and 2.

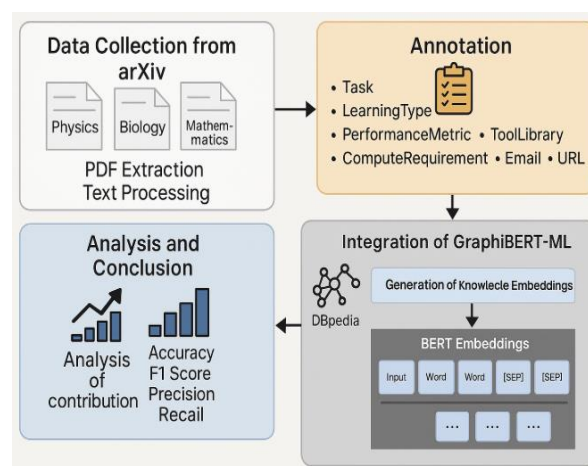


Figure 2. Schema of the Proposed Architecture

Table 1. Distribution of articles and sentences per domain

Scientific Domain	Number of Articles	Total Sentences	Avg. per Article
Biology	100	26178	261.78
Economics	100	75049	750.49
Electrical E	100	50542	505.42
Finance	100	72846	728.46
Computer S	100	85634	856.34
Mathematics	100	76635	766.35
Physics	100	63307	633.07
Statistics	100	70805	708.05
Total	800	520996	-

2.1 Data collection

To develop our datasets dedicated to exploring the application of machine learning across various academic domains, particular attention was paid to rigorous classification based on arXiv's primary domains, ranging from physics to biology and finance. This approach ensures a clear and structured division, essential for a meaningful cross-domain comparative analysis. Each article was examined through the lens of specific criteria, reflecting the nuances and diversity of machine learning approaches.

At the core of our analysis, we placed significant importance on the technical aspects of the publications. A substantial part of this annotation process concerns the nature of the machine learning tasks addressed, thus revealing the variety of technical methodologies spanning the different domains studied. In particular, we observe task types e.g., classification, regression, clustering, or reinforcement learning, contributing to an increased appreciation of the variety and creativity of technical solutions. We note that this identification exercise showcases variability and cleverness in the way these complex problems are addressed, displaying both implementation decisions and the type of algorithmic technique favored by each study.

Furthermore, this classification leads to an understanding of the aims and goals of the application of ML, and the generic difficulties each area tries to address. The separation between supervised, unsupervised, semi-supervised or reinforcement based learning models is another important dimension of our approach. This makes it possible to not only clarify the learning strategies used but, also, the theoretical and practical bases for research in each field.

The monitoring of performance by tracking the performance measures that were explicitly measured in the papers also contributes to an understanding of the effects and outcomes obtained through the application of machine learning. These data could facilitate an accurate evaluation of the method studied, comparing precision, recall, area under the curve (AUC) or other measures. Moreover, the number of results for software tools and libraries like TensorFlow [21] and PyTorch [22] found made clear, that we could potentially gauge the present popularity and tendency of implementing machine learning-based solutions by mining those

statistics. Likewise, hardware requirements (e.g., types of computation including CPU, GPU, TPU, etc., and the learning time) are important considerations, as they give us an idea of the resources for achieving the reported performance.

Together, these aspects yield an overview about the efficiency, accessibility and sustainability of machine learning techniques investigated. Further than just technical issues, our analysis covers infrastructure and implementation details as well as the software used and computational needs, also aspects related to collaboration and publication, like direct access to original materials via statistics and URLs, future exchanges and collaborations by providing the email addresses of the authors.

Table 2 provides these categories of entities and such entities used for tagging arXiv articles. it offers a short summary of each category albeit to organize and classify machine learning methods from many literatures and several academic domains.

Table 2. Labels and Descriptions Used in Annotation

Entity Type	Entity	Description
Task	classification	Indicate the type of machine learning task
	regression	
	clustering	
LearningType	supervised	Indicate the type of learning used
	unsupervised	
	semi-supervised	
	reinforcement	
PerformanceMetrics	precision	Indicate the performance metrics used to evaluate the models
	f1 score	
	recall	
	auc	
	roc	
	accuracy	
	confusion matrix	
ToolLibrary	tensorflow	Indicate the libraries or tools used in machine learning
	pytorch	
	scikit-learn	
	transformers	
	numpy	
	pandas	
	keras	
	xgboost	
	matplotlib	
ComputeRequirement	cpu	Indicate the computational requirements needed for training or executing the model
	gpu	
	tpu	
Email	Email : @	Indicate the presence of an email address
URL	URL	Indicate the presence of a URL

This extensive annotation procedure, along with our two-filter lens on academic domain and application of machine learning, enables us to provide a fine-grained exploration regarding the influence of machine learning and its evolution across diverse academic setting. It seeks to promote our understanding on which the innovative impact of machine learning in the paradigm of modern science becomes more clear, by also

featuring samples of meaningful contributions and innovations in methodology within areas of interest.

Annotations were performed manually for a detailed and contextual analysis of every publication. This approach offered a systematic analysis of the articles, based on predefined needs, such as machine learning task, learning model and performance metric. By meticulously examining each reference we were able to capture nuances and intricate details that were pertinent to a full understanding of the methodologies and aims employed. The manual approach therefore laid a strong basis for the systematic interrogation and cross-discipline comparison applied in this investigation.

2.2 Annotation strategy

For constructing these datasets, we developed a Python-based script intended for the purpose of efficiently filter articles based on two main criteria: the academic domain of arXiv and their relevance for machine learning research. This dual-criterion filtering enabled the precise selection of eight distinct datasets, each representing a specific field of study on arXiv rich in machine learning research. By limiting our search to these two simultaneous filters, the script automatically identified and downloaded 100 articles for each key domain, such as physics, finance, and biology, addressing various issues and solutions in machine learning. The choice of 100 articles per domain was made to ensure sufficiently broad and diverse coverage while keeping data management at a practical level.

We conducted a systematic and comprehensive intra-domain search using a script, covering diverse methods, applications, and technological developments unique to each domain. The post-extraction processing was performed separately for each domain and revealing the depth and diversity of machine learning research, highlighting the differences and innovations present in each area. This preliminary action is also very important from the perspective of our goal of conducting a cross-domain comparative analysis with the ultimate objective of synthesizing key insights, identifying cross-cutting trends and pinpointing research challenges and opportunities at the interface of machine learning with representative scientific fields.

The script connects to the arXiv API (Application Programming Interface) by sending HTTP (HyperText Transfer Protocol) requests and then uses BeautifulSoup to parse the XML (eXtensible Markup Language) responses. This method enables us to obtain essential details, including the articles' URLs and their distinct identifiers on arXiv. Once the articles are downloaded in Portable Document Format (PDF) via these URLs, the script saves them in a separate directory on Google Drive, making it easier to access and handle our research files.

Showing progress indicators during execution and handling errors effectively helps maintain a stable and consistent user experience. Ultimately, this script was

crucial to creating our machine learning research database by streamlining the demanding process of gathering and structuring relevant papers sourced from arXiv.

In our processing pipeline, we initiate the process by listing PDF documents located in the target directory. Next, for every PDF listed, we retrieve its text content. We clean the text data to remove unnecessary line breaks and extra spaces, to improve clarity. An important phase within this workflow is dividing the content into predetermined chunks. Such segmentation remains essential to keep these text chunks remain within set limits because of the processing and memory limitations within the Spacy library [23] may face when handling lengthy documents.

The efficiency and accuracy of our Python script in implementing these two filters for collecting and organizing articles enriches our analysis of the current machine learning landscape across various disciplines. This analytical strategy aims not only to deepen our understanding of the application and development of machine learning but also to foster innovative interdisciplinary exchanges, significantly contributing to the advancement of scientific and technological knowledge.

2.3 Pre-processing

After successfully extracting the PDF file contents, we proceed with the data processing by the identification of specific entities within the text, a crucial step for effectively structuring and analyzing the information.

These entities, which are central to our analysis, are divided into three broad categories: technical aspects of machine learning, infrastructure and implementation, and collaborative and publication aspects.

In the category of technical aspects of machine learning, we focus on the fundamental elements that characterize research and developments in the field, such as the types of machine learning (ML) tasks addressed, the learning models applied, and the performance achieved. These annotations allow us to grasp the technical scope and objectives of the studies examined.

The second category, infrastructure and implementation, gathers information about the tools and software libraries mentioned in the articles, as well as the computing requirements related to the hardware used and the training time needed. This data is essential for understanding the technological preferences and computational constraints within the machine learning research community.

Finally, the collaborative and publication aspects include the URLs of the articles, providing direct access to the original works for further exploration, and the email addresses of the authors, facilitating communication and potential future collaborations. This category highlights the importance of

collaboration and knowledge dissemination in advancing the field.

To effectively extract and structure these entities, we use a predefined extraction method, subsequently organizing all the information into a DataFrame. The entities in this DataFrame are labeled using the NER approach, and manual validation ensures the accuracy and contextual relevance of the annotations. The DataFrame is essential in our workflow, offering an organized view of the information we extract, thereby facilitating our subsequent analyses. This methodological framework ensures the quality and readiness of the data for upcoming analysis stages, representing a solid foundation for our exploration of trends and innovations in machine learning across the five specific fields of study.

2.4 GraphiBERT-ML Architecture training and evaluation: our proposed algorithm

Natural Language Processing (NLP) [24] experienced a breakthrough following the introduction of BERT (Bidirectional Encoder Representations from Transformers) [25], initiating a new direction in NLP. BERT is again a state-of-the-art model built on multiple changes in the NLP world. Similar to the Transformer's encoder module [26], BERT thoroughly handles a word sequence as input. Every layer self-attends and then passes its output through a feedforward neural network, to be passed onto the subsequent encoder.

Training multiple encoder layers introduces a significant challenge, which BERT manages by utilizing the "Masked Language Model" approach [27]. This technique involves selecting at random 10% of tokens, masking 10% of these, substituting 80% with a randomly chosen alternative, and retaining the last 15% unchanged. The model seeks to precisely recover the initially altered token (using cross-entropy loss). Consequently, it is required to preserve a contextualized distributional representation for every input token.

According to our previous article [28], BERT has demonstrated remarkable results in processing machine learning articles, particularly in named entity recognition (NER) tasks. Compared to RoBERTa [29] and DistilBERT [30], BERT has excelled in both the efficiency and accuracy of extracting named entities, confirming its superiority in these specific applications. To further improve performance in interdisciplinary scientific literature, we propose an enhanced version called GraphiBERT-ML, designed to integrate external domain knowledge into the model. This extension maintains the lightweight nature of the BERT architecture while adding a layer of contextual richness to support better generalization across academic fields.

To achieve this, we developed an algorithm that incorporates knowledge embeddings derived from textual descriptions extracted from DBpedia. These descriptions provide complementary semantic information about annotated machine learning terms in

scientific articles. Our process begins by querying DBpedia using SPARQL to retrieve these textual descriptions. When a description is found, it is converted into a simulated embedding vector; otherwise, a zero vector is assigned.

GraphiBERT-ML combines pretrained contextual representations from BERT with external semantic knowledge extracted from DBpedia to enhance named entity recognition across scientific domains. The overall process can be described in four main phases.

In the initial state, all annotated sequences and their corresponding entity labels are extracted from the dataset D . From these sequences, domain-specific terms $t \in T$ are identified, and each term is assigned an initial knowledge embedding $K(t)$ set to zero.

During Phase 1 – Knowledge Embedding Extraction, the algorithm retrieves semantic descriptions from DBpedia through SPARQL queries. For every term t , the system attempts to obtain a textual description. When a description is found, it is transformed into a knowledge embedding vector $K(t)$. If no information is available, the term retains a zero vector, ensuring consistent dimensionality across all embeddings.

In Phase 2 – Representation Enrichment, each scientific sequence $S \in D$ is tokenized using the BERT tokenizer. For every token t in the sequence, the pre-trained contextual embedding generated by BERT is concatenated with the corresponding knowledge vector $K(t)$. This produces enriched token representations that combine linguistic features with external semantic signals.

Phase 3 – Training and Evaluation uses a k -fold cross-validation strategy. For each fold, the model is trained on the enriched embeddings and evaluated on a corresponding test split. This step measures how effectively GraphiBERT-ML generalizes across domains when external knowledge is integrated.

Finally, in Phase 4 – Performance Computation, evaluation metrics including precision, recall, and F1-score are calculated for each fold. The process is repeated until all folds are processed, ensuring a robust and comprehensive assessment of the model's performance across multiple scientific disciplines.

To complement the description of the algorithm previously provided, Figure 3 illustrates the overall architecture of the proposed GraphiBERT-ML model. The process begins with the preparation of the annotated dataset, where sequences and entity labels are extracted, followed by the retrieval of external semantic knowledge from DBpedia and the generation of corresponding embedding vectors. These knowledge-based embeddings are then combined with BERT token representations to produce enriched token-level vectors. The resulting representations are subsequently used during the model training and evaluation phases,

enabling more accurate and context-aware named entity recognition across scientific domains.

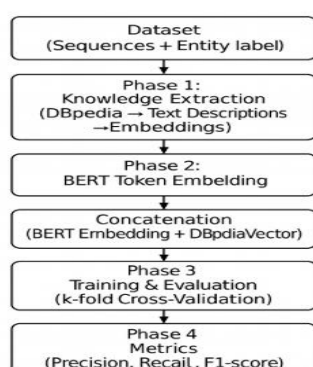


Figure 3. GraphiBERT-ML architecture

To promote reproducibility and support future research, we plan to release the dataset, preprocessing pipeline, and the full implementation of the GraphiBERT-ML model once the extended version of this work is finalized. The public package will be accompanied by comprehensive documentation, enabling researchers to reproduce the experiments and further explore knowledge-enhanced NER approaches across scientific domains.

3. Results and Discussions

In this section, the rapid proliferation of knowledge calls for methods that distill and synthesize information more effectively. Our research applies machine learning techniques to data from various arXiv domains, revealing practical insights from this rich information source.

A comparative analysis of named entity recognition (NER) tasks shows varying performance levels across disciplines, as presented in Table 3. Computer science leads with the highest metrics, including, a recall of 0.9372, F1-score of 0.9368, a precision of 0.9376, a recall of 0.9372, and an accuracy of 0.9372. This result stems the advanced state of machine learning methodologies and well-structured data in this field.

Table 3. Model Performance Metrics across Different Academic Domains

Domain	Accuracy	F1-score	Precision	Recall
Physics	0.9129	0.9115	0.9146	0.9129
Mathematics	0.8976	0.8970	0.8982	0.8976
Computer Sc	0.9372	0.9368	0.9376	0.9372
Biology	0.7967	0.7946	0.7989	0.7967
Finance	0.7887	0.7872	0.7884	0.7887
Economics	0.8413	0.8384	0.8420	0.8413
Statistics	0.8637	0.8619	0.8635	0.8637
Electrical/sys	0.8468	0.8463	0.8483	0.8468

The fields of physics and mathematics demonstrate respectable performances, although slightly lower than that of computer science. This is attributed to the semi-structured nature of their data, which supports machine learning techniques but still presents moderate variability. As these fields continue to adopt machine learning methods, further optimizations, such as

domain-specific model adaptations, could enhance NER task performance.

In contrast, the domains of biology, finance, and economics exhibit lower performance, with accuracy scores ranging from 0.7887 to 0.8413. This decline can be explained by the semantic complexity and variability of their textual data, which makes it difficult for standard machine learning models to generalize effectively. In biology, the frequent use of acronyms—such as PCR (Polymerase Chain Reaction, a widely used laboratory method) and RNA (Ribonucleic Acid)—as well as long and highly specialized gene or protein names complicates consistent entity recognition. To address this, external resources such as biomedical ontologies (e.g., MeSH, UniProt) can be used to normalize terminology. In finance, terminological ambiguity is frequent: words like equity may denote shareholder equity or company stocks, while turnover can mean revenue in some regions and staff rotation in others. Standardized financial glossaries such as IFRS (International Financial Reporting Standards) help reduce these inconsistencies. Together, these challenges highlight the need for domain-specific resources and tailored adaptations to improve NER performance in heterogeneous corpora.

Computer science achieves top performance metrics, reaching an accuracy of 0.9372, reflecting the maturity of machine learning tools and techniques in this domain. The structured and standardized datasets in computer science facilitate the application of advanced NER models, leveraging advanced model frameworks to achieve consistent and robust results. This highlights the importance of structured data and well-established methodologies in achieving superior performance in entity recognition tasks.

Electrical engineering and systems sciences, as well as statistics, present mixed results, largely due to incomplete or limited datasets. Despite this limitation, statistics demonstrates a positive trend, achieving an accuracy of 0.8637, which reflects the alignment between machine learning models and structured, quantitative data. Improving data availability and structure in these fields could unlock further potential for NER performance.

Overall, this cross-domain analysis emphasizes the critical role of data structure and domain-specific adaptations in determining the success of machine learning applications. Fields with standardized and curated data, such as computer science, achieve superior performance, whereas less structured fields, such as biology and finance, face unique challenges requiring tailored solutions. By addressing these challenges, this study provides actionable insights for optimizing NER performance and advancing interdisciplinary collaboration across academic disciplines.

To further analyze the model behavior, we conducted a qualitative error analysis. In addition, some misclassifications were observed in specific domains. In biology, certain long gene or protein names (e.g., BRCA1, TP53) are occasionally misclassified. In finance, ambiguous terms such as “equity” or “turnover” may receive incorrect labels depending on context. These errors are mainly related to semantic ambiguity and domain-specific terminology. Compared to baseline BERT, GraphiBERT-ML shows a moderate improvement in F1-score across most domains.

4. Conclusions

In this way, our work brings forward a new approach to exploring and analysing ML publications on arXiv. As described for the introduction, our purpose was to make an exhaustive –as interdisciplinary as possible– study of what trends, challenges, and opportunities which could potentially emerge due to innovation. Results and discussion validate these expectations and show evidence that our method is capable of pinpointing and annotating seminal features of publications, thus emphasizing machine learning as a booster of cross-disciplinary research.

Using the GraphiBERT-ML approach, which integrates BERT representations with the knowledge embeddings of graphs, we obtained encouraging results, especially in computer science. These results highlight the possibility of using machine learning as a tool to enable a comprehensive understanding and collaborative innovation in different research fields. To the best of our knowledge, this work represents the first attempt to integrate DBpedia-based knowledge embeddings into a transformer architecture for large-scale, cross-domain scientific NER.

In the future, we aim to apply our approach to other scientific preprint platforms to generalize our results, and to obtain a more nuanced view of the effect of the penetration of machine learning in academia. Moreover, we would like to automate the annotation process by leveraging state-of-the-art machine learning approaches. It could supplant the manual validation step which will enhance scalability and help to maintain consistency between very large datasets. Future work we will refine extraction to deal with more complex entity types and explore pre-training in domain to increase accuracy. We will also optimize the performance of the GraphiBERT-ML algorithm by exploiting other knowledge sources and fine-tuning it for specific tasks. These can include further generalizing knowledge embeddings to new subfields and exploring new training methods to improve the contextuality of BERT-based models.

In conclusion, our research contributes to academic discourse and opens new opportunities for collaboration across disciplines. By equipping researchers with effective analytical tools, we hope to support scientific progress in an era of rapid innovation. We believe these

efforts will help advance interdisciplinary research and foster meaningful knowledge exchange within the scientific community.

Acknowledgements

The authors declare no conflict of interest.

References

- [1] B. S. Lubis and A. Yus, “Mapping knowledge and research trend on technology adoption in higher education: A bibliometric analysis,” *Education and Information Technologies*, Jun. 2024, doi: [10.1007/s10639-024-12801-0](https://doi.org/10.1007/s10639-024-12801-0).
- [2] A. Sharma, Amrita, S. Chakraborty, and S. Kumar, “Named Entity Recognition in Natural Language Processing: A Systematic Review,” in *Proceedings of the Second Doctoral Symposium on Computational Intelligence*, Singapore: Springer, 2022, pp. 817–828, doi: [10.1007/978-981-16-3346-1_66](https://doi.org/10.1007/978-981-16-3346-1_66).
- [3] Cornell University Library, “arXiv.org e-Print archive,” [Online]. Available: <https://arxiv.org/>. Accessed: Jun. 20, 2025.
- [4] A. Goyal, V. Gupta, and M. Kumar, “Recent Named Entity Recognition and Classification techniques: A systematic review,” *Computer Science Review*, vol. 29, pp. 21–43, Aug. 2018, doi: [10.1016/j.cosrev.2018.06.001](https://doi.org/10.1016/j.cosrev.2018.06.001).
- [5] S. Naseer, M. M. Ghafoor, S. K. Alvi, A. Kiran, G. M. S. Rahmand, and G. Murtaza, “Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance,” *Pakistan Journal of Multidisciplinary Research*, vol. 2, no. 2, 2021.
- [6] W. Otto, M. Zloch, L. Gan, S. Karmakar, and S. Dietze, “GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Dec. 2023, pp. 8166–8176, doi: [10.18653/v1/2023.findings-emnlp.548](https://doi.org/10.18653/v1/2023.findings-emnlp.548).
- [7] O. Ovtšarenko, “Opportunities of machine learning algorithms for education,” *Discover Education*, vol. 3, no. 1, p. 209, Nov. 2024, doi: [10.1007/s44217-024-00313-5](https://doi.org/10.1007/s44217-024-00313-5).
- [8] J. Li, A. Sun, J. Han, and C. Li, “A Survey on Deep Learning for Named Entity Recognition,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: [10.1109/TKDE.2020.2981314](https://doi.org/10.1109/TKDE.2020.2981314).
- [9] S. F. Yilmaz, F. B. Mutlu, I. Balaban, and S. S. Kozat, “TMD-NER: Turkish multi-domain named entity recognition for informal texts,” *Signal, Image and Video Processing*, vol. 18, no. 3, pp. 2255–2263, Apr. 2024, doi: [10.1007/s11760-023-02898-0](https://doi.org/10.1007/s11760-023-02898-0).
- [10] H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Computer Science*, vol. 2, no. 3, p. 160, Mar. 2021, doi: [10.1007/s42979-021-00592-x](https://doi.org/10.1007/s42979-021-00592-x).
- [11] E. Harsányi et al., “Assessment of Advanced Machine and Deep Learning Approaches for Predicting CO₂ Emissions from Agricultural Lands,” *Earth Systems and Environment*, vol. 8, no. 4, pp. 1109–1125, Dec. 2024, doi: [10.1007/s41748-024-00424-x](https://doi.org/10.1007/s41748-024-00424-x).
- [12] J. Dunstan et al., “A pseudonymized corpus of occupational health narratives for clinical entity recognition in Spanish,” *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 204, Jul. 2024, doi: [10.1186/s12911-024-02609-w](https://doi.org/10.1186/s12911-024-02609-w).
- [13] Z. Liu et al., “CrossNER: Evaluating Cross-Domain Named Entity Recognition,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, pp. 13584–13592, May 2021, doi: [10.1609/aaai.v35i15.17587](https://doi.org/10.1609/aaai.v35i15.17587).
- [14] W. Shao et al., “Prompt-NER: Zero-shot Named Entity Recognition in Astronomy Literature via Large Language Models,” *arXiv preprint*, Oct. 2023, doi: [10.48550/arXiv.2310.17892](https://doi.org/10.48550/arXiv.2310.17892).
- [15] S. Wu et al., “Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in

- Nephrology,” NEJM AI, vol. 1, no. 2, Jan. 2024, doi: [10.1056/AIdbp2300092](https://doi.org/10.1056/AIdbp2300092).
- [16] J. Yeom et al., “Tc-llama 2: Fine-tuning LLM for technology and commercialization applications,” Journal of Big Data, vol. 11, no. 1, p. 100, Aug. 2024, doi: [10.1186/s40537-024-00963-0](https://doi.org/10.1186/s40537-024-00963-0).
- [17] J. Dietrich and A. Hollstein, “Performance and Reproducibility of Large Language Models in Named Entity Recognition,” Drug Safety, Dec. 2024, doi: [10.1007/s40264-024-01499-1](https://doi.org/10.1007/s40264-024-01499-1).
- [18] A. Thukral, S. Dhiman, R. Meher, and P. Bedi, “Knowledge graph enrichment from clinical narratives using NLP, NER, and biomedical ontologies,” International Journal of Information Technology, vol. 15, no. 1, pp. 53–65, Jan. 2023, doi: [10.1007/s41870-022-01145-y](https://doi.org/10.1007/s41870-022-01145-y).
- [19] S. D. Baum, “Artificial Interdisciplinarity: Artificial Intelligence for Research on Complex Societal Problems,” Philosophy & Technology, vol. 34, no. 1, pp. 45–63, Nov. 2021, doi: [10.1007/s13347-020-00416-5](https://doi.org/10.1007/s13347-020-00416-5).
- [20] P. Sun, X. Yang, X. Zhao, and Z. Wang, “An Overview of Named Entity Recognition,” in 2018 International Conference on Asian Language Processing (IALP), Nov. 2018, pp. 273–278, doi: [10.1109/IALP.2018.8629225](https://doi.org/10.1109/IALP.2018.8629225).
- [21] B. Pang, E. Nijkamp, and Y. N. Wu, “Deep Learning With TensorFlow: A Review,” Journal of Educational and Behavioral Statistics, vol. 45, no. 2, pp. 227–248, Apr. 2020, doi: [10.3102/1076998619872761](https://doi.org/10.3102/1076998619872761).
- [22] H. Dai et al., “Reveal training performance mystery between TensorFlow and PyTorch in the single GPU environment,” Science China Information Sciences, vol. 65, no. 1, p. 112103, Dec. 2021, doi: [10.1007/s11432-020-3182-1](https://doi.org/10.1007/s11432-020-3182-1).
- [23] S. Ferro, R. Giovanelli, M. Leeson, M. De Bernardin, and A. Traviglia, “A novel NLP-driven approach for enriching artefact descriptions, provenance, and entities in cultural heritage,” Neural Comput. Appl., Aug. 2025, doi: [10.1007/s00521-025-11449-2](https://doi.org/10.1007/s00521-025-11449-2).
- [24] D. Khurana, A. Koli, K. Khatte, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” Multimedia Tools and Applications, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: [10.1007/s11042-022-13428-4](https://doi.org/10.1007/s11042-022-13428-4).
- [25] S. Gorla et al., “Telugu named entity recognition using BERT,” International Journal of Data Science and Analytics, vol. 14, no. 2, pp. 127–140, Aug. 2022, doi: [10.1007/s41060-021-00305-w](https://doi.org/10.1007/s41060-021-00305-w).
- [26] R. Balamurugan, S. Mohite, and S. P. Raja, “Protein Sequence Classification Using BERT,” SN Computer Science, vol. 4, no. 5, p. 481, Jun. 2023, doi: [10.1007/s42979-023-01980-1](https://doi.org/10.1007/s42979-023-01980-1).
- [27] S. Nahali et al., “StructmRNA: A BERT-based model with dual level and conditional masking for mRNA representation,” Scientific Reports, vol. 14, no. 1, p. 26043, Oct. 2024, doi: [10.1038/s41598-024-77172-5](https://doi.org/10.1038/s41598-024-77172-5).
- [28] N. Khouya, A. Retbi, and S. Bennani, “Enriching Ontology with Named Entity Recognition (NER) Integration,” in Proceedings of ACR’24, Cham: Springer, 2024, pp. 151–159, doi: [10.1007/978-3-031-56950-0_13](https://doi.org/10.1007/978-3-031-56950-0_13).
- [29] . Chopra et al., “Roberta and BERT: Revolutionizing Mental Healthcare Through NLP,” SN Computer Science, vol. 5, no. 7, p. 889, Sep. 2024, doi: [10.1007/s42979-024-03202-8](https://doi.org/10.1007/s42979-024-03202-8).
- [30] M. Jojoa et al., “NLP applied to COVID-19 open-text opinions using DistilBERT,” AI & Society, vol. 39, no. 3, pp. 883–890, Jun. 2024, doi: [10.1007/s00146-022-01594-w](https://doi.org/10.1007/s00146-022-01594-w).