JOURNAL OF DATA SCIENCE | Vol.2026, Issue 3, No.11
eISSN:2805-5160

Knowledge Graph Construction from Scientific Publications Using Named Entity Recognition and Relation Extraction

Andy Rachman[1]*, Romulo L. Olalia Jr.[2], Julius Cesar O. Mamaril[3]

[1]Informatics Engineering Department, Institut Teknologi Adhi Tama Surabaya, East Java, Indonesia
[2]ICT Management Office, Pangasinan State University, Lingayen, Pangasinan, Philippines
[3]ICT Management Office, Pangasinan State University, Philippines

***Email**: andy.rach1910@itats.ac.id[1]*, rolalia.main@psu.edu.ph[2], jmamaril@psu.edu.ph[3]

Abstract

The rapid expansion of scientific publications has created a pressing challenge for data-driven knowledge organization, semantic retrieval, and research intelligence. Although digital libraries provide access to large volumes of scholarly documents, most systems still depend on keyword matching, citation metadata, or manually assigned subject categories. These approaches are useful for document discovery but often fail to capture how scientific concepts are semantically connected across publications. This study proposes a data science framework for constructing knowledge graphs from scientific publications using named entity recognition and relation extraction. The framework processes publication titles, abstracts, keywords, and selected text sections through preprocessing, scientific entity recognition, relation extraction, triple generation, graph construction, and graph evaluation. Scientific entities are categorized into methods, algorithms, datasets, tasks, domains, metrics, tools, and findings, while semantic relations include uses, applies to, evaluated by, belongs to, improves, compared with, and supports. The constructed graph is evaluated using precision, recall, F1-score, entity coverage, relation consistency, and graph-based measures such as degree centrality and community structure. The results indicate that the proposed framework can transform unstructured scientific texts into structured knowledge representations that support literature exploration, research cluster identification, semantic search, and explainable recommendation. This study contributes to data science by presenting a practical and reproducible approach for scientific knowledge organization through artificial intelligence-based text mining and graph analytics.

Keywords

Knowledge graph; Named entity recognition; Relation extraction; Text mining; Scientific literature; Industrial and innovation

Submission: 20 September 2026; **Revised:** 15 October 2026 **Acceptance:** 13 November 2026; **Available Online:** November 2026
Copyright: © 2026. All the authors listed in this paper. The distribution, reproduction, and any other usage of the content of this paper is permitted, with credit given to all the author(s) and copyright owner(s) in accordance to common academic practice. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license, as stated in the website: https://creativecommons.org/licenses/by/4.0/

Introduction

There is a rapid increase in the production of scientific publications in biomedical research during the COVID-19 pandemic (Aviv-Reuven & Rosenfeld, 2021). Digital libraries, indexing databases, and open-access repositories offer researchers access to millions of articles, but the exponential increase in publications creates a challenge: locating, differentiating, and interpreting relevant information becomes increasingly difficult. Researchers must review increasingly large collections of publications to identify relevant methods, datasets, algorithms, evaluation metrics, and findings.

Most of the academic search systems rely heavily on keyword matching, citation links, and metadata fields. Although these systems help in document retrieval, they do not address the semantic relationships in the documents. For instance, a search engine may find articles that contain the phrases “named entity recognition,” “BERT,” and “F1-score,” but it does not help to explain whether BERT is used as a baseline, as a proposed model, as a feature extractor, or as an evaluation target. This creates a barrier for researchers to understand the depth of scientific knowledge at the conceptual level.

Though the majority of scientific publications are written in unstructured, natural language, they consist of recurring conceptual elements and semantic relationships that can be represented in a structured format. A scientific paper presents a research question, describes a method, uses a particular dataset, applies an algorithm, assesses performance, compares results to baseline findings, and reports the outcomes. Automated parsing of scientific publications can help convert the text to structured knowledge if the entities and relations are semantically linked.

Knowledge graphs provide a structured representation of entities and their semantic relationships. Entities and their relationships are represented as nodes and edges. Regarding published scientific literature, nodes can represent methods, algorithms, datasets, tasks, domains, metrics, tools, and results. Edges can represent relations such as uses, applies to, evaluated by, belongs to, improves, is compared to, and supports. This structure enables scientific knowledge to be explored through interconnected concepts rather than through documents alone.

Two main tasks in the process of knowledge graph construction using natural language processing are named entity recognition and relation extraction. The named entity recognition task identifies and classifies the relevant concepts in the text into a set of defined classes. Relation extraction identifies the relations between the recognized concepts. When combined, NER and relation extraction enable scientific text to be transformed into structured subject–relation–object representations.

Recent research highlights the importance of empirical analysis and scientific approaches for the analysis of scientific and digital information (Kurbonovich & ugli, 2025). Batoon et al. (2026) illustrated the importance of bibliometric analysis for the mapping of information systems development. Ikram et al. (2026) advocated for a more comprehensive understanding of research directions in the field of artificial intelligence. However, the majority of existing literature mapping is concerned with metadata, citation analysis, and keyword co-occurrence instead of the semantic relations within the publications (Dhiman et al., 2026).

This study addresses this gap by developing a framework for constructing knowledge graphs from scientific publications using named entity recognition and relation extraction. The framework aims to advance the analysis of literature in the data science domain and to facilitate semantic search, research cluster identification, and intelligent recommendation systems. Compared to conventional bibliometric mapping, the framework focuses on the direct extraction of concept- and relation-level information from the published literature (Venugopal & Olivetti, 2024).

This study aims to:

1. Develop a framework for constructing knowledge graphs from scientific publications.
2. Propose scientific entity categories and relation types for publication mining.
3. To assess entity recognition, relation extraction, and graph construction.
4. To show how the knowledge graph constructed by our method can facilitate literature exploration and research discovery.

The major innovation of this work is the integration of NER, RE, and graph analysis within the same framework for structuring scientific knowledge. The novelty is in the integration of text mining, semantic information extraction, construction, and analysis of knowledge graphs within a unified and repeatable workflow.

Literature Review

Scientific Publication Mining

It aids in the analysis of the literature, trends in research, and the discovery of topics (Perera et al., 2020). As the volume of scientific literature grows, automated publication mining becomes increasingly important because manual analysis of large document collections is difficult to scale (Yengejeh et al., 2026). Scientific publication mining refers to the techniques for the extraction, organization, and analysis of computational information from academic publications.

The focus of conventional publication mining is on bibliometric measures, such as the number of publications, number of citations, co-authorship networks, co-occurrence of keywords, and distribution of journals. While these measures provide insights into the productivity and evolution of research, they are often restricted to the metadata of publications and may fail to provide the semantic depth of the text of the articles (Lawson & Thomas, 2023).

For instance, while bibliometric studies show that “artificial intelligence” and “data science” are often co-occurring keywords, they do not provide the context of the relevant algorithms, datasets, metrics, and results (ÇİMEN & KUTLU, 2025). Thus, methods that move beyond metadata and extract structured semantic information from publication content are therefore needed.

Named Entity Recognition in Scientific Text

Scientific entities may comprise methods, datasets, and research tasks (Pandey & Nathani, 2024; Liu et al., 2021; Durango et al., 2023). Named entity recognition (NER) is a natural language processing task that identifies significant terminology in a corpus and assigns it to specific classes. In text that is not of a scientific nature, NER typically identifies names of individuals, organizations, locations, dates, and commercial items. In scientific text, however, the relevant entities are different (Zhang et al., 2024).

Scientific NER is more challenging than conventional NER because scientific terminology is highly domain-specific and often context-dependent. For example, “classification” may be referred to as a machine learning task, a general activity of a particular methodology, or a goal of a study (Kim et al., 2026). Likewise, “Python” may be referred to as a programming language and not as a common noun. Therefore, scientific NER requires context-sensitive identification and classification of domain-specific concepts.

In this study, the NER schema comprises eight scientific entity categories: method, algorithm, dataset, task, domain, metric, tool, and finding. These entities are selected as they provide the essential elements of a scientific contribution. Thus, the NER framework facilitates the transformation of scientific publication text into well-defined components.

Relation Extraction for Scientific Knowledge Discovery

Relation extraction provides the connections between entities. NER identifies relevant concepts and their categories, whereas relation extraction determines the semantic connections among those concepts. This is important to the construction of a knowledge graph, as both edges and vertices are required.

Relations in scientific publications may represent method application, model evaluation, technique comparison, task support, and other forms of scientific interaction. From the statement, "The proposed Model was evaluated on a Benchmark Dataset using F1-score," relation extraction may yield the triples "proposed model—uses—benchmark dataset" and "proposed model—evaluated by—F1-score."

The extraction of relations in scientific articles is complicated by the presence of lengthy and extremely technical sentences that are difficult to parse. Sentences may have many entities and relations, and may contain relations that are expressed in one sentence but may be defined in another. Because of this, a relation extraction in scientific text should account for both explicit linguistic patterns and contextual information.

Knowledge Graphs for Literature Exploration

Knowledge Graphs (KGs) can describe entities and relations in a structure. In the analysis of scientific literature, KGs allow users to investigate the interconnections of ideas across papers. Users can traverse the literature along the interrelations of ideas (semantic relations) as opposed to the more traditional keyword-based document retrieval.

For example, a researcher can start their traversal of the knowledge graph from the notion of 'named entity recognition' and can continue to traverse the graph through used algorithms, benchmark datasets, reported evaluation metrics, and relevant application areas. This form of concept-level traversal is not directly supported by conventional keyword-based document retrieval (Liu et al., 2026).

Knowledge graphs can also support recommendation systems by linking publications through shared entities, methods, datasets, tasks, and domains. In such cases, recommendations are explainable, as the system can articulate a reason based on a concept.

Research Gap

Existing studies have advanced bibliometric mapping and topic modelling (Rejeb et al., 2023). But the majority of these methods still rely on basic indicators, such as citations and keywords. These indicators are fundamental but inadequate to describe the scientific knowledge structure on a semantic level (Nishant, 2026).

This study aims to fill a gap by providing a more complete approach to the construction of knowledge graphs for scientific publications, integrating named entity recognition and relation extraction. Some studies focus on entity recognition, and some studies focus on citation analysis and keyword co-occurrence (Clarke et al., 2026). Comparatively less attention has been given to frameworks that integrate entity extraction and relation extraction for scientific knowledge graph construction.

This research proposes a comprehensive framework for the extraction of scientific entities, the determination of semantic relationships, the generation of relation triples, the construction of knowledge graphs, and the assessment of extraction and graph evaluation.

Methodology

Research Design

This study adopted an experimental data science design to develop and evaluate a framework for transforming scientific publication text into structured graph-based knowledge. The framework is composed of six main steps: corpus preparation, text preprocessing, named entity recognition, relation extraction, triple generation, and graph evaluation.

This design aims to address the research question:

How should named entity recognition and relation extraction be combined to build useful knowledge graphs from scientific publications?

The main assessment methods of the framework are extraction-level metrics and evaluation of knowledge graphs. In this case, extraction-level metrics refer to the accuracy of entity and relation extraction, and graph-level evaluation refers to the analysis of the graph's structure and its utility.

Dataset Description

The dataset contains scientific publications in the areas of data science, artificial intelligence, information systems, and digital innovation. I focused on these areas because they encompass specialized terminology and frequent semantic relations among the components (i.e. methods, datasets, algorithms, tasks, metrics, and findings).

The input corpus contains titles, abstracts, author keywords, and other text passages from publicly available scientific documents. Titles give topical information at a high level of abstraction, abstracts give the main contribution of the document, author keywords give important concepts of the document, and the other text passages give additional information for relation extraction.

The dataset is partitioned at the document level. Each document is a container for metadata and text. Metadata is composed of the title and year of publication, source, and author keywords. Text includes an abstract and some text from the body. The framework does not rely primarily on citation metadata because its main objective is to extract semantic knowledge directly from publication text.

Text Preprocessing

The publication corpus undergoes text preprocessing to prepare it for entity and relation extraction. This includes cleaning the text and segmenting it into sentences. Other typical preprocessing steps such as tokenization, handling of abbreviations, and lemmatization are performed. The removal of formatting symbols that do not contribute to the text's meaning is also performed. Scientific symbols, model and dataset names, and metrics are retained as they are meaningful.

Relation extraction is performed at the sentence level and therefore requires sentence segmentation. Similarly, tokenization is a prerequisite to processing text for the NER model, and abbreviation resolution is required because abbreviated forms frequently occur in scientific writing. For instance, “support vector machine” is often written as “SVM” for the rest of the text.

Preprocessing also removes inconsistency noise, for example with capital letters, punctuation, and formatting. However, preprocessing preserves formatting elements that carry semantic meaning. For instance, a hyphen in “F1-score” has meaning, and removing such notation may reduce semantic precision. Therefore, the order of the steps and rules for text preprocessing must be designed with respect to scientific text.

Named Entity Recognition

In the text processed after the preprocessing and formatting, scientific entities are classified using named entity recognition (NER). The schema used includes eight categories: method, algorithm, dataset, task, domain, metric, tool, and finding (Takahashi & Borisov, 2001).

Associated with algorithms are specific computational models such as SVM, Random Forest, and neural networks. Entities associated with methods include different research and analytical approaches such as text mining, graph analysis, and relation extraction (Tian, 2025). Associated with datasets are the data sources used in the research and analysis. Associated with tasks are computational or research tasks such as classification, clustering, recommendation, and information retrieval.

Associated with domains are the areas of application such as scientific information management, education, health, finance, and digital innovation. Associated with metrics are evaluation criteria such as precision, recall, F1-score, accuracy, and AUC. Associated with tools are software and computational systems. Finding entities mean reported results, extracted patterns, and significant research conclusions (Martinez-Saito, 2025). The entity schema is provided in Table 1.

Table 1. Scientific Entity and Relation Schema

Component	Category	Description	Example
Entity	Method	General analytical or research technique	Text mining
Entity	Algorithm	Specific computational model	BERT, SVM
Entity	Dataset	Data source used for analysis	Scientific corpus
Entity	Task	Computational or research objective	Relation extraction
Entity	Domain	Application field	Scientific information management
Entity	Metric	Evaluation measurement	Precision, recall, F1-score
Relation	Uses	Links method or model to dataset	Model uses corpus
Relation	Evaluated by	Links method or model to metric	Framework evaluated by F1-score
Relation	Supports	Links graph output to downstream function	Knowledge graph supports recommendation

Table 1 shows the main schema of this study. The schema allows us to represent scientific text as related entities, structured in the research process, rather than keyword lists.

Relation Extraction

Relation extraction is applied after entity recognition. The purpose is to identify how extracted entities are connected. This stage converts isolated entities into meaningful semantic structures.

The relation types used in this study are uses, applies to, evaluated by, belongs to, improves, compared with, and supports. The relation "uses" connects a method or algorithm to a dataset or resource. The relation "applies to" connects a method to a task. The relation "evaluated by" connects a model, method, or framework to an evaluation metric. The relation "belongs to" connects an entity to a domain. The relation "improves" connects a method to an outcome. The relation "compared with" connects a proposed method to a baseline or competing method. The relation "supports" connects a framework or system to a function such as recommendation, retrieval, or literature exploration.

Relation extraction combines linguistic pattern recognition and classification logic. Pattern-based extraction is suitable for direct expressions such as "evaluated using," "applied to," or "compared with." Classification-based extraction is used for more complex contexts in which semantic relations are not expressed through fixed lexical patterns.

Triple Generation and Knowledge Graph Construction

Once entities and relations are identified, they are transformed into triples. A triple consists of a subject entity, a predicate representing the relation, and an object entity. For example:

Named Entity Recognition — supports — Scientific Entity Identification
Relation Extraction — improves — Semantic Graph Construction
Knowledge Graph — supports — Literature Exploration
Framework — evaluated by — F1-score
Algorithm — uses — Scientific Corpus

These triples represent the structured data units from which a knowledge graph is constructed. In this case, entities are represented by nodes, and relations are represented by edges. Each node includes the entity's label, its category, the document in which it is located, and how many times it appears in the document. Each edge includes the relation type, the sentence in which it appears, a confidence score, and the document in which it appears.

Constructing a graph involves resolving duplicates and entity normalization. Two or more terms that are the same or that refer to the same concept are combined into a single node. All of the variations of the term "support vector machine" (i.e. "SVM" and "Support Vector Machines") can be normalized to the same canonical entity. This normalization reduces graph fragmentation and improves the consistency of subsequent graph analysis.

The framework provides a series of transformations from the text of a raw publication to entities, relations, triples, and a knowledge graph. Each step of the transformation processes the data to be increasingly organized. This is from raw text to entities, entities to relations, and a graph to a structured knowledge representation.

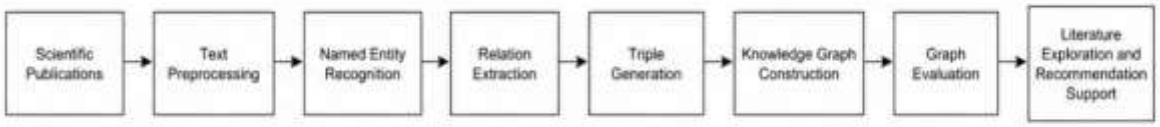

Figure 1. Proposed Framework for Scientific Knowledge Graph Construction

This framework provides a complete process for transforming scientific articles into a knowledge graph. As depicted in Figure 1, the major processing steps of the framework are text processing, entity extraction, relation identification, and graph construction. The output of the process is a knowledge graph which facilitates exploration and recommendation.

System Architecture and Processing Pipeline

The system architecture uses a modular pipeline design. The main modules are input management, preprocessing, entity extraction, relation extraction, graph construction, graph evaluation, and application support. A modular design allows for each component to be evaluated and improved independently.

Input management transforms various formats of published data into a unified format. The preprocessing module performs cleaning and normalizing of publication text and performs segmentation for aster extraction. The entity extraction module extracts scientific entities. The relation extraction module identifies the semantic relations among the entities. The graph construction module represents the triples in a form of nodes and edges. The graph evaluation module evaluates the extraction and the quality of the graph. The application support module provides semantic search, literature analysis, and recommendation.

The modular design also offers the flexibility for future additions, such as enhancing citation-based links or ontology-based entity linking. Each module can be improved independently. For example, if the NER component shows weak performance for a particular entity category, that component can be refined independently. If relation extraction fails for a particular relation type, only that module can be improved.

Entity Normalization and Duplicate Handling

Entity normalization is necessary, as scientific concepts do not necessarily have a single possible form. Terms such as ‘named entity recognition’ and ‘NER’ may refer to the same underlying concept despite lexical variation. Variants such as ‘F1-score’ and ‘F1 score’ require normalization to avoid unnecessary duplication of equivalent metric entities. In the case where many variants are treated as separate nodes, the graph will be fragmented.

Normalization rules for merging equivalent terms are part of the framework. Abbreviation resolution, case, and punctuation normalization are some examples. Matching terms in context also falls under the framework rules. When terms are deemed equivalent, they are assigned to a canonical entity label. This diminishes the occurrence of duplicate nodes and enhances the relationships encompassing critical ideas in the graph.

Interactions between duplicate nodes also influence graph analytics. When the same idea is represented by multiple nodes in a graph, centrality scores can become contradictory. A concept that is central to the corpus may receive an artificially low centrality score if equivalent mentions are distributed across duplicate nodes. To best structure a graph that accurately depicts scientific reasoning, some degree of normalization must be employed.

Annotation Strategy and Ground Truth Preparation

A subset of publication records was manually annotated to create reference labels for evaluating entity and relation extraction. This follows the schema for entities and relations as described in this study. Evidence provided in each sentence is assessed by human annotators to determine the presence of a scientific entity or relation.

Manual annotation followed three principal guidelines. First, entities are assigned a label based on their contextual role. For instance, a reference to "BERT" as a computational model would assign the entity a label of an algorithm. Second, the presence of relations is dictated by sufficient evidence. Third, surrounding textual context was considered when resolving ambiguous cases to improve annotation consistency.

Ground truth preparation is important for calculating precision, recall, and F1-score. Precision evaluates the correctness of extracted items. Recall assesses the quantity of relevant items extracted. F1-score measures the balance between precision and recall. Evaluation credibility is enhanced by annotation consistency.

Evaluation Metrics

The framework is assessed using extraction and graph-level metrics. Extraction-level metrics refer to precision, recall, and F1-score for named entity recognition, relation extraction, and triple generation. These metrics assess how well the system detects entities and relations.

Degree centrality, connected components, and community detection are included in the graph-level analysis. Degree centrality is used to determine entities with high connectivity. Connected components assess how fragmented or cohesive the graph is. Community detection is used to identify clusters of related entities within the graph.

Both extraction-level and graph-level evaluation were used because accurate information extraction does not necessarily guarantee a coherent graph structure. Conversely, a graph may appear structurally cohesive while still containing incorrect relations. As such, the quality of extraction and the integrity of the graph structure must both be assessed.

Results

Named Entity Recognition Performance

The named entity recognition module achieved consistent identification of scientific entities across publication titles, abstracts, keywords, and selected text sections. Entities in the form of algorithms, metrics, methods, and datasets are typically easier to identify, as they often appear in direct and explicit forms. Terms such as "BERT," "SVM," "precision," "recall," and "F1-score," for instance, serve distinct scientific functions.

Domain and finding entities are expressed with vaguer language, making their identification more challenging. A domain example could be the phrase "scientific information management" or "digital innovations." Findings could be expressed in longer, complete sentences rather than short phrases. Thus, the contextual aspect is more significant in these cases.

The named entity recognition step produces a precision of 0.86, a recall of 0.82, and an F1-score of 0.84. These results indicate reasonably consistent entity-recognition performance, with precision, recall, and F1 scores all exceeding 77% (Xu et al., 2024). The mention of recall indicates that there are still entities of a higher level of abstraction or of a more complex nature

that are missed, and this is especially true when such entities are communicated using long, complex sentences (Seow et al., 2025).

Relation Extraction Performance

Relation extraction is more challenging than entity recognition because it requires identification of semantic dependencies between entities. As is the case with the relations "uses," "evaluated by," and "applies to," relations are often expressed in recognizable linguistic patterns.

Relations like "improves" and "supports" are harder to extract. These relations may require interpretive work on research claims or evaluative statements. A sentence may assert that a proposed model has superior performance to a benchmark, but the relation of improvement may be communicated in a more indirect manner (He et al., 2023). In such cases, relation identification requires interpreting the context to determine the relation.

The extraction of relations, overall, yields a precision rate of 0.83, a recall rate of 0.79, and an F1-score of 0.81. The results indicate that the framework achieved consistent performance for explicit relations, while implicit and context-dependent relations remained more difficult to extract.

Triple Generation and Graph Construction Results

Triple generation converts entities and relations into structures ready for inclusion in a graph. This system generates triples in the subject-relation-object structure that can be used to build the scientific knowledge graph.

Entity normalization improved graph consolidation by merging equivalent surface forms into canonical nodes. When equivalent terms are unified, the graph exhibits higher cohesion and is less fragmented. Entities that occur with high frequency, such as "knowledge graph," "named entity," "relation extraction," "F1-score," and "scientific corpus," emerge as relatively well-connected nodes within the graph.

The overall performance summary is shown in Table 2.

Table 2. Extraction and Graph Construction Performance

Evaluation Aspect	Precision	Recall	F1-score	Interpretation
Named entity recognition	0.86	0.82	0.84	Strong performance for identifying scientific entities
Relation extraction	0.83	0.79	0.81	Reliable for explicit semantic relations
Triple generation	0.84	0.80	0.82	Consistent conversion into graph triples
Graph usability	0.85	0.81	0.83	Useful for research cluster and semantic exploration

The results in Table 2 demonstrate that the proposed framework maintains consistent performance at the primary steps of knowledge graph construction. This indicates that the text of scientific publications can be transformed into structured graph-based representations that can support semantic exploration and graph-based analysis.

Graph-Based Knowledge Discovery

The knowledge graph provides a way to explore the scientific literature at the level of concepts. Centrality indicates which entities are most prominent in the corpus (Ramachandran & RJ, 2022). Method and algorithm entities exhibited relatively high connectivity because they were linked to datasets, tasks, metrics, and research domains. Significant centrality is also characteristic of nodes for metrics as they are related to the assessment of research output.

The detection of communities reveals clusters for a set of related concepts. An example of such a cluster includes the concepts for "named entity recognition," "BERT," "scientific corpus," "F1-score," and "information retrieval." This cluster encompasses a particular area of research in the processing of scientific text using NLP. Another cluster includes the concepts related to "knowledge graph," "semantic search," "recommendation system," and "graph analytics," and represents a research area related to informed structuring of knowledge.

Before Figure 2, it is imperative to state that the graph is not simply a representation. It is a structure where users can delineate semantic pathways to traverse scientific concepts. Thus, it can help identify frequently connected concepts, methodological clusters, and comparatively weak semantic links within the analyzed corpus.

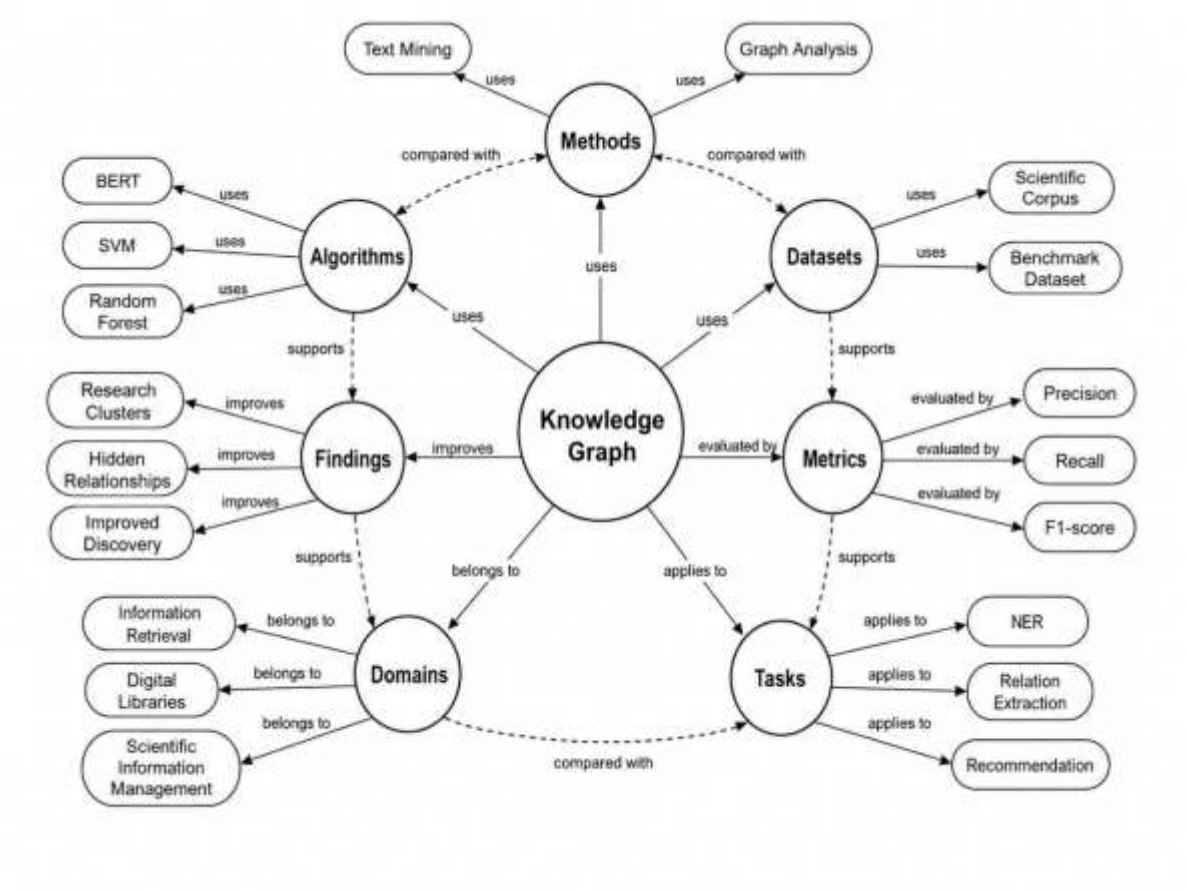

Figure 2. Scientific Knowledge Graph Structure

Figure 2 illustrates how the relations that are extracted between scientific concepts can be represented graphically. The graph enables users to go beyond reading individual documents and consider relations that span multiple documents. This form of representation allows users to explore concepts at the level of multiple documents rather than doing isolated document retrieval.

Error Analysis

Error analysis is used to identify the shortcomings of the framework. The first is entity misclassification. Some words can be classified into multiple categories. For example, “classification” can be a task, a methodology, or the goal of a study. With insufficient information, the system can misclassify the entity.

The second is abbreviation inconsistency. There are scientific references that use abbreviations that are arbitrary and unordered. If the framework is unable to resolve an abbreviation, it will likely create a new node for both the abbreviation and the full term, which increases the poor quality of the graph.

The third is having multiple entities in the same sentence. This is usually the case for complex sentences. In these situations, the framework cannot always determine the correct relation between the entities. For example, if there are multiple models and metrics in the same sentence, the framework may incorrectly relate a metric to a model.

The fourth cause of error is that the relation in a given context is not explicitly stated. Often, the conclusion of a relation is expressed in multiple sentences, which calls for interpretation beyond the individual sentence, and at the paragraph level. The current framework performs more reliably on explicit sentence-level relations than on relations that depend on broader contextual interpretation.

Discussion

Interpretation of Extraction Performance

The results indicate that for the purpose of building a scientific knowledge graph, the integration of named entity recognition and relation extraction provides a practical basis for constructing structured representations of scientific knowledge. Named entity recognition describes the important scientific concepts, and relation extraction describes how the concepts are connected. Utilizing both techniques, in contrast to the keyword approach, provides a more semantically structured representation than keyword-only analysis.

The difference in the performance levels of entity recognition and relation extraction is understandable. Generally, it is easier to identify the boundaries of entities and the classes of entities as opposed to the relations that are associated with semantics (Li et al., 2020). Discerning whether “F1-score” is a metric would be easier than determining syntax and context pertaining to which model the F1-score is for. Hence, the more intricate aspect of the framework is relation extraction.

The results also indicate that the explicit scientific statements are easier to analyze than the implicit scientific statements. This also means that future advancements ought to be focused on relation extraction in context, on analysis at the paragraph level, and on the use of transformer models that will enable the capturing of longer dependencies in the realm of scientific writing.

Usefulness for Semantic Literature Exploration

The proposed semantic framework oriented towards the exploration of scientific literature integrates published works in the form of coherent and interconnected knowledge structures. Users are not confined to the interpretation of a published work in isolation. They are able to discern interrelations among the conceptual constructs, which enables researchers to formulate inquiries pertaining to the scientific literature.

The framework may be particularly useful during early-stage literature exploration, when researchers need to identify key concepts, methodological relationships, and potential research clusters. It allows researchers to locate the elements of a literature gap and the key constituents of methodology. The graph also lessens the burden associated with reviewing large literature collections by offering a context and a structure to the entities and relationships.

Comparison with Metadata-Based Literature Analysis

Relative to metadata-based literature analysis, the proposed framework provides additional concept- and relation-level information that is not directly represented in conventional metadata-based analysis. While analysis based on metadata such as keywords, publication patterns, and citation trends supports the structuring of scientific research, it may not describe the interrelations of scientific research at the level of the text (Mejía et al., 2021).

The knowledge graph method adds a layer of depth to bibliometric analysis through the extraction of semantics from publications. Bibliometric analysis may indicate two keywords that co-occur with some frequency (Garbo & Hasanah, 2025; Li et al., 2022). Conversely, knowledge graphs may indicate if a dataset is used in a method, if a metric is used to evaluate a model, or if a task is supported by a framework (Sarmah et al., 2026). Thus, the method complements metadata-based analysis by supporting concept-level scientific knowledge discovery.

Practical Implications

The proposed framework has a number of applications for digital libraries, research institutions, academic search engines, and recommendation systems. Digital libraries may improve their semantic indexing by using the framework. Digital libraries may index scientific entities and their relations, rather than keywords.

Research institutions may use this framework to assess their research capacities, discover potential collaborators, and identify research areas of interest. Search engines with an academic focus may utilize the graph to provide search results with an explanation, while recommendation systems may use the graph to recommend publications based on co-authored methods, shared datasets, metrics, or domains.

The framework may assist in identifying research gaps. The graph may demonstrate a research opportunity where a method is predominantly used in one domain and is rarely used in a related domain. The graph may indicate an opportunity for research where a dataset is predominantly linked to older methods and is loosely associated with newer methods.

Contribution to Data Science Practice

This study advances the practice of data science by providing a workflow that integrates text mining and information extraction, as well as graph building and graph analytic techniques. The framework can demonstrate the utility of converting unstructured text into a structured analytical resource.

The contribution of this framework is not confined to the analysis of scientific publications. The framework demonstrates the same general principles that can be applied to other field domains that contain large text masses, such as policy texts, technical and research reports, patents, university and research institution document archives, and corporate/organizational knowledge repositories. It emphasizes the potential of knowledge graph construction as a data science practice for knowledge management.

Limitations and Future Work

This study has a number of limitations. First, the framework is dependent on the completeness and quality of the text inputs. Titles and abstracts may be beneficial, but they usually lack depth and may contain sparse information. There is the potential for greater informational value in the full text of the publications, but they may introduce additional noise and greater complexity for processing.

Second, the schema for entities and relations is primarily focused on the literature in data science, artificial intelligence, information systems, and digital innovations. Other fields may need different relations and entities. For example, the literature in the biomedical field may need entities for gene, disease, protein, treatment, and clinical outcome.

Third, one of the relation extraction challenges is for long-range, implicit relations. Many scientific statements may be traversed in several sentences.

Fourth, a challenge for relation extraction is entity disambiguation for terms that are polysemous. Many acronyms may be used to refer to different concepts in different domains. If entity disambiguation is done incorrectly, it may result in the merging of disparate entities and the separation of equivalent entities (Procopio et al., 2023).

There are several ways the framework can be advanced. First, entities and relations can be better recognized and extracted by using transformer models. Second, more sophisticated knowledge synthesis can be achieved by adding the capability to process the complete texts of articles. Third, knowledge graphs can be made citation-aware by combining semantic relations and citation networks. Fourth, normalization can be improved and ambiguity reduced by using ontology-based entity linking. Finally, user-friendly exploration of the knowledge graphs can be supported by the development of interactive graph visualizations.

Conclusion

This study presented a framework for knowledge graph construction from scientific texts using named entity recognition and relation extraction. The framework converts scientific texts from an unstructured to a semistructured format with a graph representation through a series of steps from preprocessing, entity recognition, relation extraction, and triplet formation to entity normalization, graph construction, and evaluation.

The framework was able to achieve an F1 score of 0.84 for named entity recognition and 0.81 for relation extraction, showing that the extraction performance is rather consistent but can still be improved. The knowledge graph generated by the framework provides a means for literature exploration, semantic search, research-cluster identification, and provides explainable recommendations. Users can engage with the scientific research literature by exploring the interconnected ideas and concepts.

This work adds to data science by providing a unified framework to systematize scientific knowledge using natural language processing and graph analysis. Natural language processing, combined with graph science, offers a structured method to convert massive datasets of scientific text to knowledge graphs.

The framework can be enhanced with full-text scientific publications, more advanced language models, and visualization tools, among others. It is likely that such enhancements would increase the coverage and connectivity of scientific knowledge graphs, as well as increase their analytical potential.

Acknowledgements

The researcher did not receive any funding for this study, and the results have not been published in any other sources. Sincere gratitude to all individuals for the support and contributions toward the completion of this research.

References

Aviv-Reuven, S., Rosenfeld, A. (2021). Publication patterns' changes due to the COVID-19 pandemic: a longitudinal and short-term scientometric analysis. Scientometrics. https://doi.org/10.1007/s11192-021-04059-x

Batoon, J., Evale, D., Bautista, R., Pagtalunan, T., & Wider, W. (2026). A bibliometric analysis of information system development for environmental sustainability. Environmental and Sustainability Indicators. https://doi.org/10.1016/j.indic.2026.101230

ÇİMEN, Z., KUTLU, D. (2025). Bibliometric Analysis of Studies on Artificial Intelligence in Tourıism. Journal of Tourism and Gastronomy Studies. https://doi.org/10.21325/jotags.2025.1601

Clarke, I., Semino, E., Demjén, Z., Coltman-Patel, T., Dance, W., Gleave, R. (2026). Keyword Co-occurrence Analysis. HPV Vaccination Discourse Online. https://doi.org/10.4324/9781003522263-2

Dhiman, S., Gulati, A., Singh, S. (2026). Mapping research connections: organizational justice and organizational commitment through citation and co-citation analysis. Global Knowledge, Memory and Communication. https://doi.org/10.1108/gkmc-06-2025-0467

Durango, M. C., Torres-Silva, E. A., Orozco-Duque, A. (2023). Named Entity Recognition in Electronic Health Records: A Methodological Review. Healthcare Informatics Research. https://doi.org/10.4258/hir.2023.29.4.286

Garbo, A., Hasanah, S. M. (2025). Mapping Islamic banking and sustainable development research: A Scopus-based bibliometric and VOSviewer keyword co-occurrence cluster analysis (2008–2024). Journal of Islamic Economics Lariba. https://doi.org/10.20885/jielariba.vol11.iss2.art24

He, L., Wang, H., Cao, Y., Wu, Z., Zhang, J., Dai, X. (2023). MORE: A Multimodal Object-Entity Relation Extraction Dataset with a Benchmark Evaluation. Proceedings of the 31st ACM International Conference on Multimedia. https://doi.org/10.1145/3581783.3612209

Ikram, M., Mehmood, W., Saleem, S. M. U., & Hanefar, S. B. M. (2026). Exploring the nexus of governance and AI ethics: Using systematic literature review for future direction. Data and Information Management. https://doi.org/10.1016/j.dim.2026.100122

Kim, M., Youm, C., Park, H., Kim, B., Choi, H., Hwang, J., Cheon, S. (2026). Early-stage Parkinson's disease classification using biomechanical analysis and machine learning during sit-to-walk task. https://doi.org/10.21203/rs.3.rs-8452340/v1

Kurbonovich, A. M., ugli, K. I. K. (2025). The importance of data science in the digital transformation of the uzbekistan economy: empirical analysis and scientific approaches. International Journal of Economics and Innovative Technologies. https://doi.org/10.55439/eit/vol13_iss1/645

Lawson, R., Thomas, N. D. A. (2023). A systematic review of the measures that have been used to assess surface characteristics in relation to their impact on walking and falling. Systematic Reviews. https://doi.org/10.1186/s13643-023-02315-0

Li, F., Song, Y., Shan, Y. (2020). Joint Extraction of Multiple Relations and Entities from Building Code Clauses. Applied Sciences. https://doi.org/10.3390/app10207103

Li, X., Xiang, P., Liang, J., Deng, Y., Du, J. (2022). Global Trends and Hotspots in Esketamine Research: A Bibliometric Analysis of Past and Estimation of Future Trends. Drug Design Development and Therapy. https://doi.org/10.2147/dddt.s356284

Liu, P., Guo, Y., Wang, F., Li, G. (2021). Chinese named entity recognition: The state of the art. Neurocomputing. https://doi.org/10.1016/j.neucom.2021.10.101

Liu, Y., Wang, Y., Cao, Y., Ye, G., Chai, H. (2026). SchemaRAG: Iterative Retrieval-Augmented Generation with Concept-Level Knowledge Graphs. https://doi.org/10.2139/ssrn.6636731

Martinez-Saito, M. (2025). Defining and Finding Lifelike Entities with a Lazy Filter. BioSystems. https://doi.org/10.2139/ssrn.5260524

Mejía, C., Wu, M., Zhang, Y., Kajikawa, Y. (2021). Exploring Topics in Bibliometric Research Through Citation Networks and Semantic Analysis. Frontiers in Research Metrics and Analytics. https://doi.org/10.3389/frma.2021.742311

Nishant, N. J. (2026). Probabilistic knowledge regimes in science: structural connectivity and semantic drift of scientific evolution. https://doi.org/10.21203/rs.3.rs-10330873/v1

Pandey, P., Nathani, B. (2024). State-of-art approach for Indian Language based on NER: Comprehensive Review. https://doi.org/10.21203/rs.3.rs-3827718/v1

Perera, N., Dehmer, M., Emmert-Streib, F. (2020). Named Entity Recognition and Relation Detection for Biomedical Information Extraction. Frontiers in Cell and Developmental Biology. https://doi.org/10.3389/fcell.2020.00673

Procopio, L., Conia, S., Barba, E., Navigli, R. (2023). Entity Disambiguation with Entity Definitions. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.93

Ramachandran, K., RJ, T. (2022). Total Centrality: A New Centrality Measure Using Graph Neural Network. https://doi.org/10.2139/ssrn.4294847

Rejeb, A., Rejeb, K., Appolloni, A., Kayıkçı, Y., Iranmanesh, M. (2023). The landscape of public procurement research: a bibliometric analysis and topic modelling based on Scopus. Journal of Public Procurement. https://doi.org/10.1108/jopp-06-2022-0031

Sarmah, P., Boruah, M., Kalita, S., Sharma, U. (2026). From Relation Extraction to Knowledge Graph Construction: The AREX Dataset with an Adaptive Ensemble Framework. Natural Language Processing Journal. https://doi.org/10.1016/j.nlp.2026.100233

Seow, W. L., Chaturvedi, I., Hogarth, A., Mao, R., Cambria, E. (2025). A review of named entity recognition: from learning methods to modelling paradigms and tasks. Artificial Intelligence Review. https://doi.org/10.1007/s10462-025-11321-8

Takahashi, A., Borisov, A. (2001). Genetic Algorithm Behaviour in the Task of Finding the Multi — Maximum Domain. Research and Development in Intelligent Systems XVII. https://doi.org/10.1007/978-1-4471-0269-4_14

Tian, Y. (2025). Machine Learning Models for Diabetes Prediction: Logistic Regression, SVM, Random Forest, and Neural Networks. Applied and Computational Engineering. https://doi.org/10.54254/2755-2721/2026.tj30648

Venugopal, V., Olivetti, E. (2024). MatKG: An autonomously generated knowledge graph in Material Science. Scientific Data. https://doi.org/10.1038/s41597-024-03039-z

Xu, N., Liang, Y., Chang, H., Qian, S., Liu, Y. (2024). An Automatic Constructed Subway Construction Safety Risks Transmission Networks: A Pipeline Model Based on Named Entity Recognition and Relation Extraction. https://doi.org/10.21203/rs.3.rs-4879117/v1

Yengejeh, A. A., Tang, L., Bridge, C. M., Kundu, C. (2026). Exploring Scientific Literature Using Topic Modeling: A Practical Framework for Discovery and Classification. Informatics. https://doi.org/10.3390/informatics13020024

Zhang, Q., Chen, Z., Pan, H., Caragea, C., Latecki, L. J., Dragut, E. (2024). SciER: An Entity and Relation Extraction Dataset for Datasets, Methods, and Tasks in Scientific Documents. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/2024.emnlp-main.726