Journal of Information Systems Engineering and Management
2025, 10(41s)

e-ISSN: 2468-4376

https://www.jisem-journal.com/

Research Article

Intelligent Semantic Search for Academic Journals Using AI and NLP Techniques

[1,3] Shireen Fathi Malo, *[2]Adel Al-zebari

[1, 2]Department of Information Technology, Technical College of Informatics-Akre, Akre University for Applied Sciences, Akre, Duhok, Kurdistan region, Iraq. adel.hasan@auas.edu.krd, shireen.fathi@auas.edu.krd

[3]Department of Information Technology, Technical College of Informatics-Akre, Duhok Polytechnic University, Duhok, Kurdistan region, Iraq. shireen.fathi@auas.edu.krd

ARTICLE INFO

Received: 17 Nov 2024

Accepted: 26 Dec 2024

ABSTRACT

The exponential growth of academic literature has rendered traditional keyword-based search engines increasingly inadequate for scholars seeking contextually relevant research. This study presents the design and implementation of an intelligent semantic search engine tailored for academic journals, integrating state-of-the-art Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques. The proposed system leverages sentence transformer models (all-mpnet-base-v2) for semantic embeddings, enabling vector-based similarity searches, alongside spaCy for tokenization and entity recognition to enhance syntactic understanding. An ontology-based matching mechanism further aligns user queries with domain-specific research topics, while fuzzy matching and regular expressions improve error tolerance and numeric filtering (e.g., CiteScore, Impact Factor). The system architecture combines these NLP layers with Elasticsearch's hybrid search capabilities to process and rank peer-reviewed journal metadata sourced from Scopus and DOAJ. A modular FastAPI-based backend ensures scalability and responsiveness, while a lightweight frontend interface facilitates interactive user input. This research contributes a novel hybrid framework that unites neural semantic models with structured query construction, addressing limitations in current scholarly search systems. The study also introduces a benchmark methodology for evaluating semantic search performance in academic contexts, with implications for enhancing research efficiency, interdisciplinary discovery, and access to high-impact literature.

KEYWORDS: Semantic Search, Natural Language Processing (NLP), Artificial Intelligence (AI), Sentence Transformers, Academic Journal Retrieval

1. INTRODUCTION

In the digital age, the exponential growth of academic literature has made efficient discovery of relevant research a critical challenge for scholars [1]. Traditional search engines, which rely on keyword matching and Boolean operators, often fall short in addressing the nuanced demands of researchers [2]. These systems struggle with semantic ambiguity, fail to interpret context, and prioritize lexical overlap over conceptual relevance, leading to incomplete or irrelevant results [3]. As the volume of interdisciplinary research expands, the limitations of conventional search tools—such as their inability

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

to process natural language queries or map relationships between complex ideas—have become increasingly apparent [4] [5].

The problem lies in bridging the gap between user intent and computational interpretation. Researchers frequently formulate queries as descriptive questions (e.g., "How does climate change affect biodiversity in tropical forests?"), yet existing tools lack the capacity to infer meaning, contextualize terminology, or prioritize results based on semantic similarity [6], [7]. This disconnect underscores the need for a more intelligent search paradigm—one that leverages advances in artificial intelligence (AI) to decode intent, analyze context, and deliver precise, domain-specific insights.

This study aims to address this gap by developing an AI-powered semantic search engine tailored for academic journals. Unlike generic search systems, the proposed framework integrates three core innovations: (1) AI-driven text embeddings to map contextual relationships between queries and documents, (2) ontology-based matching to align concepts with domain-specific knowledge hierarchies, and (3) fuzzy search techniques to accommodate typographical errors and terminological variations. By focusing exclusively on peer-reviewed journals—rather than individual papers or preprints—the system targets a critical yet underserved niche, streamlining access to high-impact, vetted research.

The scope of this work is deliberately bounded to ensure feasibility and depth. It excludes non-journal publications (e.g., conference proceedings, theses) and prioritizes scalability within structured academic databases. This constraint allows the model to refine its understanding of journal-specific discourse patterns and citation networks, enhancing its ability to surface contextually relevant results.

The contributions of this research are threefold. First, it introduces a hybrid architecture combining transformer-based language models with lightweight ontologies, enabling both broad semantic understanding and domain-aware precision. Second, it pioneers a fuzzy search algorithm optimized for academic jargon, addressing the variability in scientific terminology. Finally, the study provides a benchmark dataset for evaluating semantic search performance in academic contexts, filling a gap in existing research infrastructure. By reimagining how scholars interact with literature, this work advances the frontier of AI-driven knowledge discovery, offering a tool that not only accelerates research but also fosters interdisciplinary connections and innovation.

2. BACKGROUND THEORY:

In order to design and implement an intelligent semantic search engine tailored for academic journals, it is essential to ground the system's architecture and functionality in established theoretical foundations. The fields of Information Retrieval (IR), Natural Language Processing (NLP), and Artificial Intelligence (AI) provide the conceptual and technical basis for understanding how textual data can be processed, represented, and queried effectively. Traditional search methods, primarily reliant on lexical keyword matching, often fail to capture the nuanced semantics of user intent, particularly in scholarly research where domain-specific terminology and complex query structures are prevalent. Consequently, recent advancements in transformer-based models, ontology-driven query expansion, and hybrid retrieval frameworks have enabled more context-aware and semantically meaningful search capabilities.

This section provides an overview of the core theoretical concepts that underpin the proposed system. It begins by examining classical information retrieval models and their evolution toward semantic search. It then introduces key NLP techniques such as tokenization, embedding generation, and entity recognition, which play a central role in query understanding. Furthermore, the section explores the application of domain ontologies, fuzzy string matching, and regular expressions for query refinement and data extraction. Finally, it addresses the role of Elasticsearch as a hybrid search engine capable of supporting both vector-based semantic similarity and structured Boolean filtering. Collectively, these

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

foundational theories inform the design choices and algorithmic components implemented in the system, bridging the gap between user queries and relevant academic content.

2.1. Information retrieval systems:

Traditional information retrieval (IR) systems initially relied on Boolean models, which allowed users to search using logical operators (AND, OR, NOT) to match exact keywords in documents. While effective for simple queries, Boolean retrieval lacked ranking capabilities, often returning overly broad or narrow results. The vector space model (VSM) addressed this by representing documents and queries as vectors in a high-dimensional space, ranking results based on cosine similarity. A key advancement was TF-IDF (Term Frequency-Inverse Document Frequency), which weighted terms by their importance in a document relative to their frequency across a corpus. Combined with inverted indexing a data structure mapping terms to their occurrences TF-IDF enabled efficient and scalable retrieval [8].

Despite these improvements, keyword-based systems struggled in scholarly contexts due to their reliance on exact term matching, ignoring synonyms, related concepts, and semantic meaning. For instance, a search for "machine learning" might miss relevant papers using terms like "statistical learning" or "neural networks." Additionally, keyword models could not interpret user intent or contextual nuances, leading to low recall or precision. These limitations spurred the development of modern semantic search techniques, leveraging natural language processing (NLP) and embedding (e.g., BERT) to understand query-document relationships beyond lexical matching [9].

2.2. Fundamentals of Neural Language Processing:

The NLP pipeline begins with tokenization, the process of splitting text into meaningful units (e.g., words, subwords), followed by part-of-speech (POS) tagging, which assigns grammatical categories (e.g., noun, verb) to tokens, aiding in syntactic analysis. Named entity recognition (NER) then identifies and classifies real-world entities (e.g., persons, organizations) within the text. These tasks operate at different linguistic levels: lexical processing (word-level analysis), syntactic processing (sentence structure), and semantic processing (meaning extraction). Early rule-based systems relied on handcrafted linguistic rules, but modern NLP leverages statistical and machine learning models, such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs), to improve accuracy [10].

Libraries like spaCy streamline NLP workflows by providing pre-trained models for tokenization, POS tagging, and NER, enabling efficient text processing without extensive manual feature engineering. Unlike traditional approaches, spaCy integrates deep learning (e.g., transformer-based models) to enhance contextual understanding, making it useful for applications like information extraction and question answering. While lexical and syntactic processing form the foundation, semantic analysis (e.g., word embeddings, dependency parsing) bridges the gap between raw text and machine comprehension, allowing NLP systems to infer intent and relationships [11]. However, challenges remain in ambiguity resolution and cross-lingual generalization, driving advancements in large language models (LLMs) like BERT and GPT.

2.3. Text embedding and transformer models:

Text embeddings are numerical representations of words or sentences that capture semantic meaning in a dense vector space. Early models like Word2Vec [12] used shallow neural networks to generate static word embeddings, mapping semantically similar words (e.g., "king" and "queen") to nearby vectors. However, these embeddings lacked context sensitivity words like "bank" (financial vs. river) had the same representation. The introduction of BERT (Devlin et al., 2019) revolutionized NLP by using transformer architectures to produce contextualized embeddings, where word meanings adjust based on surrounding text. Sentence-BERT (SBERT) [13] further improved efficiency by fine-tuning

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

BERT to generate fixed-length sentence embeddings, enabling semantic similarity comparisons without expensive pairwise computations.

A key application of embeddings is K-Nearest Neighbors (KNN) search in vector space, where semantically similar texts are retrieved based on cosine or Euclidean distance. Models like all-mpnet-base-v2, a variant of SBERT, excel in this task due to their ability to generate high-quality sentence embeddings optimized for semantic search. Unlike traditional keyword matching, KNN over embeddings captures paraphrases and conceptual relationships, making it ideal for tasks like recommendation systems and scholarly literature retrieval.

2.4. Ontologies and sematic web:

In computer science, an ontology is a formal framework that defines concepts, relationships, and constraints within a specific domain, enabling machines to interpret and reason about data meaningfully. Unlike simple taxonomies, ontologies use structured vocabularies (e.g., classes, properties, instances) to model knowledge in a machine-readable format, often employing standards like OWL (Web Ontology Language) and RDF (Resource Description Framework). In semantic search, ontologies facilitate query expansion by linking user queries to related concepts—for example, a search for "cardiac arrest" could automatically include "myocardial infarction" if defined as a synonym or subclass in a medical ontology [14]. JSON-based ontologies (e.g., Schema.org) further simplify integration with web applications, bridging unstructured text and structured knowledge graphs.

Ontologies enhance topic alignment by mapping unstructured text to predefined taxonomies. For instance, in scholarly search, a domain ontology might align "deep learning" with broader terms like "artificial intelligence" or related techniques like "convolutional neural networks," improving recall and precision. Projects like DBpedia and WordNet demonstrate how ontologies standardize terminology across datasets, enabling interoperability in the Semantic Web [15]. By embedding ontological reasoning into search systems, queries can leverage hierarchical (e.g., "mammal" → "dog") or associative (e.g., "author" → "publication") relationships, transforming keyword matches into context-aware retrievals.

2.5. Elasticsearch and Hybrid Search Engines:

Elasticsearch is a distributed, scalable search engine built on Apache Lucene, designed for fast full-text and structured data retrieval. Its architecture organizes data into indices (logical partitions), which are further divided into shards (physical subsets) to enable horizontal scaling and parallel processing. Each index has a mapping that defines the schema, including field types (e.g., text, keyword, dense_vector) and analysis rules (e.g., tokenizers, filters). Elasticsearch excels at full-text search using inverted indices and scoring algorithms like BM25, which ranks documents based on term frequency and relevance [16]. With the rise of neural search, Elasticsearch integrated vector search capabilities, allowing dense embeddings (e.g., from BERT or SBERT) to be indexed and queried alongside traditional text, enabling hybrid semantic and keyword retrieval.

Indexing is critical to both performance and accuracy in Elasticsearch. For full-text search, inverted indices enable efficient term lookup, while for vector search, approximate nearest neighbor (ANN) algorithms like HNSW (Hierarchical Navigable Small World) optimize high-dimensional similarity searches. By combining these approaches, Elasticsearch supports hybrid search, where keyword matches and semantic similarities are jointly scored to improve result quality [17]. For example, a query for "machine learning applications" might retrieve documents with exact keyword matches (lexical) alongside conceptually related papers (semantic). This flexibility makes Elasticsearch a cornerstone of modern search infrastructure, balancing speed, scalability, and relevance.

3. LITERATURE REVIEW:

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The paper [18] "Harnessing AI for Enhanced Searching in Digital Libraries: Transforming Research Practices" explores the integration of AI technologies, particularly Natural Language Processing (NLP) and Machine Learning (ML), to improve search functionalities in digital libraries. Traditional search systems often struggle with vague or ambiguous user queries, leading to irrelevant results. This study highlights how AI can enhance query comprehension, contextual understanding, and information retrieval accuracy, ultimately improving the user experience. The research reviews various query formulation methods—automated, semi-automated, and manual—to optimize searches. The study concludes that AI-powered search engines provide more relevant results, reduce user frustration, and transform research practices by enabling efficient and intuitive access to academic resources. The paper's key contributions include analyzing the limitations of conventional search methods, demonstrating AI's role in refining search accuracy, and proposing adaptive query refinement techniques. Practical implications emphasize enhanced user experiences, better retrieval efficiency, and improved research workflows, making AI-driven search tools essential for the future of digital libraries.

The authors in [19] introduces an AI-enhanced workflow for bibliometric analysis, addressing limitations in traditional keyword-based methods. The study integrates advanced AI techniques, including vector databases, Sentence Transformers, Gaussian Mixture Models (GMM), Retrieval Agents, and Large Language Models (LLMs), to enable semantic search, topic ranking, and customized literature characterization. A pilot study analyzing 223 urban science-related articles from *Nature Communications* demonstrates the system's effectiveness in extracting insightful summary statistics about research quality, scope, and characteristics. The results highlight the AI-enhanced workflow's capability to uncover deeper insights beyond conventional search methods, positioning AI as a powerful tool for research evaluation. The methodology involves vectorized document representation, dense text embedding, topic clustering, and retrieval-augmented generation for contextual search, significantly improving the accuracy and relevance of bibliometric analysis. The study's key contributions include developing an AI-driven bibliometric workflow, introducing a novel approach to literature characterization, and demonstrating its application in urban research. The practical implications extend to researchers, policymakers, and academic institutions, offering enhanced research evaluation, improved knowledge retrieval, and data-driven decision-making. The study concludes that AI-powered bibliometric analysis provides a transformative approach to understanding research trends and impact, with potential applications across multiple disciplines.

In [20] the authors addresses the escalating challenge researchers face in navigating the expanding corpus of academic literature. It introduces CyLit, an innovative framework that leverages Natural Language Processing (NLP) to automate the retrieval, summarization, and clustering of scholarly works within the cyber risk domain. The study highlights CyLit's effectiveness in providing context-specific resources and tracking emerging trends in this rapidly evolving field. By comparing CyLit's literature categorization with traditional survey papers and outputs from models like ChatGPT, the authors demonstrate its unique insights and enhanced efficiency in academic literature searches. This advancement underscores the potential of NLP techniques to revolutionize how researchers discover, analyze, and utilize academic resources, fostering progress across various knowledge domains.

In [21] The authors explores the application of Machine Learning (ML) and Natural Language Processing (NLP) in building semantic search systems, focusing on a virtual legal assistant. Traditional keyword-based search methods often fail to capture the complexity of legal language, leading to irrelevant or incomplete results. By leveraging NLP techniques such as tokenization, named entity recognition, and semantic analysis, along with ML algorithms for classification and prediction, the system can better understand the intent behind user queries and the contextual meaning of legal texts. The study demonstrates that ML and NLP-based semantic search significantly enhance accuracy, efficiency, and accessibility compared to conventional methods. The results indicate improved user

satisfaction and reduced research time for legal professionals. The paper's contributions include the development of a system that provides precise and relevant legal information, facilitating informed decision-making while increasing accessibility for both legal experts and the general public. The practical implications are significant, as such systems can streamline legal workflows, enhance legal research, and improve public access to legal knowledge.

The authors [22] introduced the AI Research Navigator, a platform designed to help researchers manage the rapidly growing body of AI literature by integrating keyword search with neural retrieval techniques. It enables users to explore research at different textual granularities and leverages a domain-specific Knowledge Graph. The system employs Sentence-BERT and SciBERT for document embedding, named entity recognition for linking concepts, and various AI-driven tools like question answering, expert search, and personalized recommendations. The introduction highlights the increasing difficulty of navigating AI research due to the surge in publications, emphasizing the limitations of general academic search engines. The platform aims to enhance research efficiency through semantic search, analytics, and collaboration tools. While the paper does not present quantitative results, it discusses how these features improve search effectiveness and knowledge organization. The conclusions recognize challenges like factual accuracy and conflicting opinions, suggesting further research in these areas. Practically, the platform helps AI researchers stay updated, identify key experts, and accelerate discovery through advanced retrieval and recommendation techniques.

The paper in [23] "Semantic Embedding-Based Recommender System for Research Paper Discovery and Subject Area Prediction" introduces a novel approach to enhance the discovery of research papers and accurately predict their subject areas by leveraging semantic embedding. The authors employ advanced natural language processing techniques to generate semantic embedding that capture the contextual nuances of research papers, facilitating more precise recommendations and subject classifications. The study's findings indicate that incorporating semantic embedding significantly improves the performance of recommender systems in the academic domain, offering more relevant paper suggestions and accurate subject area predictions. This advancement addresses the limitations of traditional keyword-based methods by considering the deeper semantic content of documents. The practical implications of this research are substantial, as it aids researchers in efficiently navigating vast academic repositories, thereby accelerating literature reviews and fostering interdisciplinary collaborations. The paper contributes to the field by integrating semantic analysis into recommendation algorithms, demonstrating the effectiveness of embedding techniques in understanding complex textual data within scholarly contexts.

The paper [24] provides a comprehensive analysis of incorporating AI and machine learning into scalable search systems. The authors explore theoretical foundations and practical implementations, focusing on advanced ranking algorithms, natural language processing (NLP) for query understanding, and optimized distributed architectures. Experiments on a large-scale dataset comprising 100 million web pages and 1 million real-world queries demonstrated significant improvements: a 15% increase in Normalized Discounted Cumulative Gain (NDCG) for complex queries and a 12% improvement in Mean Reciprocal Rank (MRR) for navigational queries, compared to traditional keyword-based approaches. The study addresses challenges in maintaining system scalability and performance, including data synchronization, real-time model updates, and resource management in distributed environments. Emerging trends such as graph neural networks and multimodal search capabilities are discussed, alongside ethical considerations and data privacy concerns. The findings offer valuable insights for developing next-generation search platforms capable of handling increasing complexity and volume of digital information while ensuring responsible AI integration.

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In [25] authors addresses the escalating challenge of manual literature reviews due to the exponential growth of research articles by developing a system that automates this process using Natural Language Processing (NLP) techniques and Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs). The study evaluates three methodologies: a frequency-based approach utilizing spaCy, a transformer-based model using Simple T5, and an LLM-based approach employing GPT-3.5-turbo. Using the SciTLDR dataset, the effectiveness of these methods was assessed through ROUGE scores, revealing that GPT-3.5-turbo achieved the highest ROUGE-1 score of 0.364, followed by the transformer model, with spaCy ranking last. The research concludes that LLMs, particularly GPT-3.5-turbo, significantly enhance the automation of literature reviews, suggesting their potential to streamline the research process. A notable contribution of this work is the development of a graphical user interface based on the LLM, facilitating practical application for researchers. The practical implications include reducing the time and effort required for literature reviews, allowing researchers to focus on critical analysis and synthesis, thereby improving efficiency in academic research.

The paper "Intelligent System for Research Article Recommendation: A Knowledge Graph-based Approach" addresses the challenges of cold start and data sparsity in recommendation algorithms by integrating knowledge graph technology to enhance intelligent literature recommendations. The authors categorize existing recommendation techniques into embedding-based, path-based, and propagation-based methods, analyzing how each extracts data from entities and connections within knowledge graphs. They introduce the KGAT-CI model, which effectively utilizes collaborative and knowledge-aware information, leading to improvements in AUC metrics by 2.9%, 1.6%, and 1.2% across three datasets compared to state-of-the-art baselines. The study concludes that incorporating knowledge graphs can significantly enhance the accuracy, diversity, and interpretability of current recommendation algorithms, thereby improving user satisfaction and the overall effectiveness of intelligent recommendation systems [26].

In [27] Rs4rs has been introduced as a web application designed to enhance the efficiency of researchers in the field of Recommender Systems by providing precise and comprehensive semantic searches for recent publications from top conferences and journals. Traditional scholarly search engines often return broad results, making it challenging to locate high-quality, relevant papers. Rs4rs addresses this issue by allowing users to input their topics of interest and receive tailored lists of pertinent papers, utilizing semantic search techniques that understand the context and meaning behind queries, thereby capturing relevant papers regardless of variations in wording. The tool significantly enhances research efficiency and accuracy, benefiting the research community and the public by facilitating access to high-quality, pertinent academic resources in the field of Recommender Systems. The platform's user-friendly interface and focus on top-tier venues ensure that researchers can efficiently stay updated with the latest advancements without the need for manual searches across multiple sources. This approach not only saves time but also supports the advancement of research by making it easier to identify trends and key developments in the field.

Year	Main Contribution	Problem	Methodology	Algorithm Used	Research Field	Results	Data Used
2024	Enhancing searching in digital libraries with AI	Improving research practices in digital libraries	AI-based search enhancements in digital libraries	AI technologies and search optimization techniques	Digital Libraries	Not specified	Not specified

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2024	Automating bibliometric analysis using sentence transformers	Improving bibliometric analysis for urban research	Using sentence transformers and RAG for semantic search	Sentence transformers, RAG	Bibliometrics, Urban Research	Improved semantic and contextual search capabilities	Not specified
2024	Developing an NLP-powered repository for cyber risk literature	Improving search in academic papers on cyber risk	NLP-powered search and organization	NLP techniques	Cyber Risk, NLP	Improved organization and search of academic papers	Cyber risk literature
2023	Semantic search for legal assistant systems	Developing a semantic search for legal information	Building a virtual legal assistant with machine learning	Machine Learning, NLP	Legal Technology	Enhanced semantic search capabilities	Legal documents
2020	New neural search and insights platform for AI research	Organizing AI research and enhancing navigation	Design of a new neural search platform	Neural search algorithms	AI Research, Information Retrieval	Improved navigation and organization of AI research	AI research publications
2024	Recommender system for research paper discovery	Improving research paper discovery and prediction	Using semantic embedding for research paper recommendation	Semantic embedding algorithms	Recommender Systems, Academic Research	Improved paper discovery	Research papers
2024	Knowledge graph-based approach for article recommendation	Enhancing article recommendation using knowledge graphs	Cold start and data sparsity in recommendation algorithms	Collaborative filtering and knowledge-aware techniques	Research Article Recommendation	Improved AUC metrics, 2.9% increase	Urban Research datasets
2024	Automated literature review using NLP and LLM	Automating literature review process	Streamlining literature review with NLP and LLM-based retrieval	NLP techniques, LLM-based retrieval	Literature Review	Improved efficiency in literature review	Research papers
2024	Rs4rs for recent publications in recommendation systems	Finding recent publications from top venues	Searching for recent publications in recommendation systems	Semantic search techniques	Recommendation Systems	Not specified	Not specified

The provided research papers collectively explore advancements in leveraging Artificial Intelligence (AI) and Natural Language Processing (NLP) to enhance various aspects of academic research, particularly in literature search, analysis, and recommendation systems. Shamsitdinova et al. (2024) discuss how AI technologies improve search functionalities in digital libraries, addressing challenges

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

like query formulation and information retrieval. Xu et al. (2024) propose a method combining Sentence Transformers and Retrieval-Augmented Generation (RAG) to automate bibliometric analysis, facilitating semantic and contextual search for urban research literature. Zhang et al. (2024) present CyLit, an NLP-powered repository and search engine designed to organize and search academic papers, focusing on cyber risk literature. Samardzhiev and Nisheva-Pavlova (2023) explore the application of machine learning and NLP in developing semantic search systems, exemplified by a virtual legal assistant. Fadaee et al. (2020) introduce a neural search and insights platform aimed at navigating and organizing AI research. Patil et al. (2024) develop a semantic embedding-based recommender system to enhance research paper discovery and predict subject areas. Biyyala et al. (2024) discuss integrating machine learning with search technology to create high-scalability AI-powered search systems. Ali et al. (2024) investigate automating literature reviews using NLP techniques and LLM-based RAG. Anil et al. (2024) propose an intelligent system for research article recommendation based on knowledge graph approaches. Wijaya et al. (2024) present Rs4rs, a tool for semantically finding recent publications from top recommendation system venues. Collectively, these studies highlight the transformative role of AI and NLP in advancing research methodologies, improving information retrieval, and facilitating personalized recommendations across diverse academic disciplines.

4. METHODOLOGY

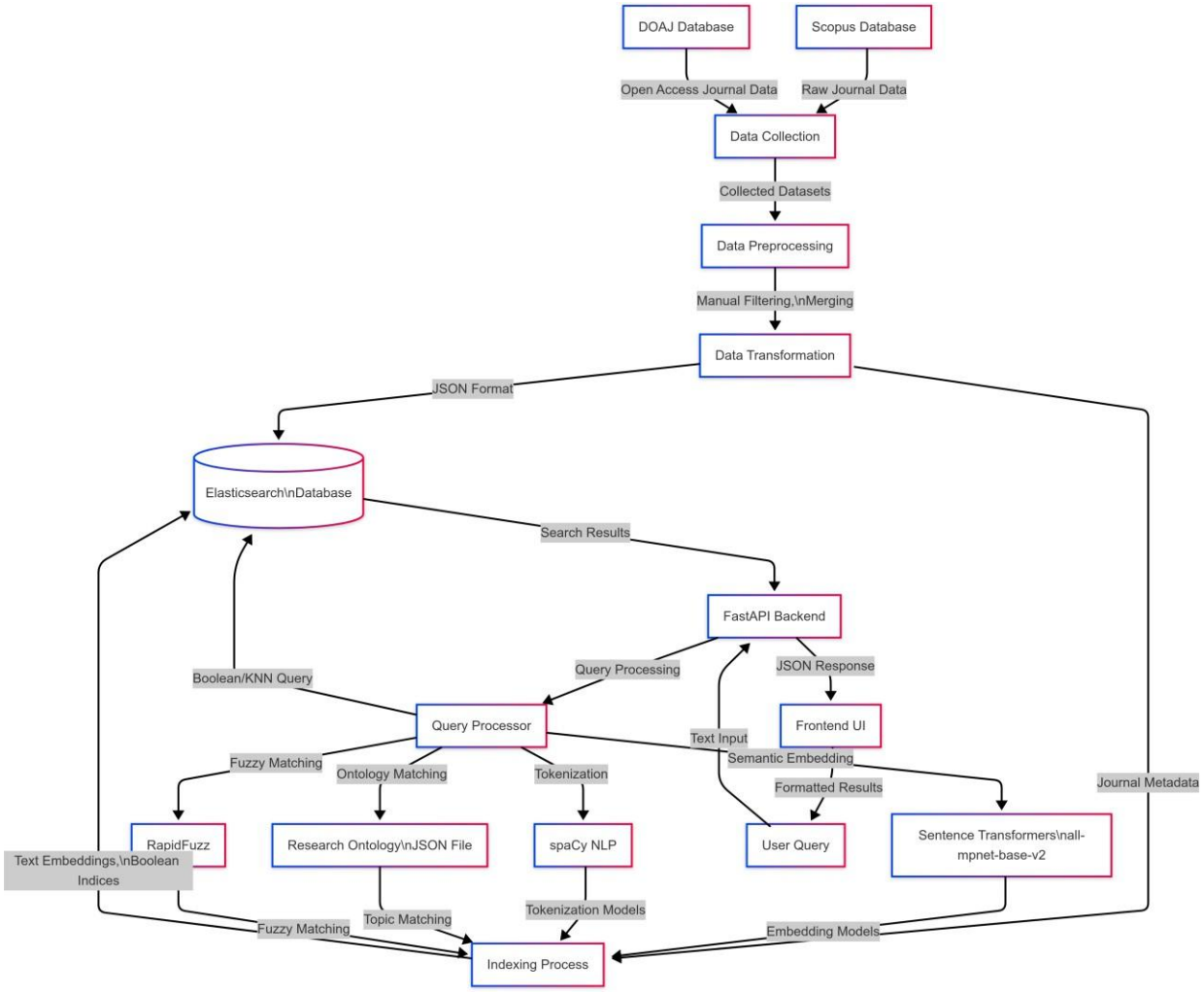
The methodology underpinning this research is designed to address the challenges of building an intelligent, scalable semantic search engine for academic journals. This section outlines the systematic approach to system development, integrating modern AI technologies, robust data engineering practices, and domain-specific optimizations. The framework is structured to balance computational efficiency with semantic depth, ensuring that the system can interpret natural language queries while maintaining the precision required for academic research.

4.1 System Design & Architecture

The system adopts a modular architecture, decoupling the frontend interface from the backend processing pipeline to ensure scalability. The backend is built using FastAPI, a high-performance Python framework optimized for asynchronous request handling, enabling rapid processing of concurrent search queries. Data storage and retrieval are managed by Elasticsearch, a distributed search engine that supports hybrid search workflows combining semantic embeddings with traditional Boolean queries. The frontend, developed with HTML, JavaScript, and CSS, provides an intuitive interface for researchers to input natural language queries and visualize ranked results. FastAPI, a modern Python web framework, is designed to optimize performance and scalability, making it particularly well-suited for handling search queries in resource-intensive applications like semantic academic search engines.

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



As shown in figure 1, The proposed methodology integrates data collection, preprocessing, indexing, and retrieval to develop an intelligent journal search system. First, journal metadata is gathered from Scopus and DOAJ databases, followed by data preprocessing and transformation, where raw data is cleaned, merged, and formatted into JSON for structured storage. The processed data is then indexed into an Elasticsearch database, where text embeddings, Boolean indices, and keyword mappings are generated to enhance search efficiency.

To support advanced querying, the system incorporates NLP and machine learning models. Sentence Transformers (all-mpnet-base-v2) generate semantic embeddings, spaCy NLP performs tokenization, and RapidFuzz enables fuzzy matching. Additionally, a research ontology JSON file ensures topic-based query refinement. When a user submits a query through the FastAPI backend, the query processor applies semantic search, Boolean filtering, and fuzzy matching before retrieving results from Elasticsearch. The results are formatted and displayed via the frontend UI, ensuring an efficient and context-aware search experience.

4.2 Data Collection and Preprocessing

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Journal metadata was aggregated from two primary sources: Scopus (for high-impact journals with citation metrics) and the Directory of Open Access Journals (DOAJ) (to include open-access publications). Data fields such as titles, abstracts, keywords, ISSNs, and citation scores (e.g., CiteScore, Impact Factor) were extracted via APIs and manually curated to remove duplicates, non-English entries, and incomplete records. The cleaned datasets were merged, normalized, and transformed into a standardized JSON schema to ensure compatibility with Elasticsearch's document-oriented storage model.

4.3 Data Storage & Indexing

Elasticsearch indices were configured to store journal documents with fields optimized for both keyword and semantic search. Text embedding, generated using the all-mpnet-base-v2 Sentence Transformer model, were indexed alongside raw text to enable k-nearest neighbors (KNN) similarity searches. Boolean query support was retained for exact field matches (e.g., ISSN, publication year). Indexing significantly accelerated retrieval times by precomputing embeddings and organizing data into inverted indexes, reducing latency during query execution.

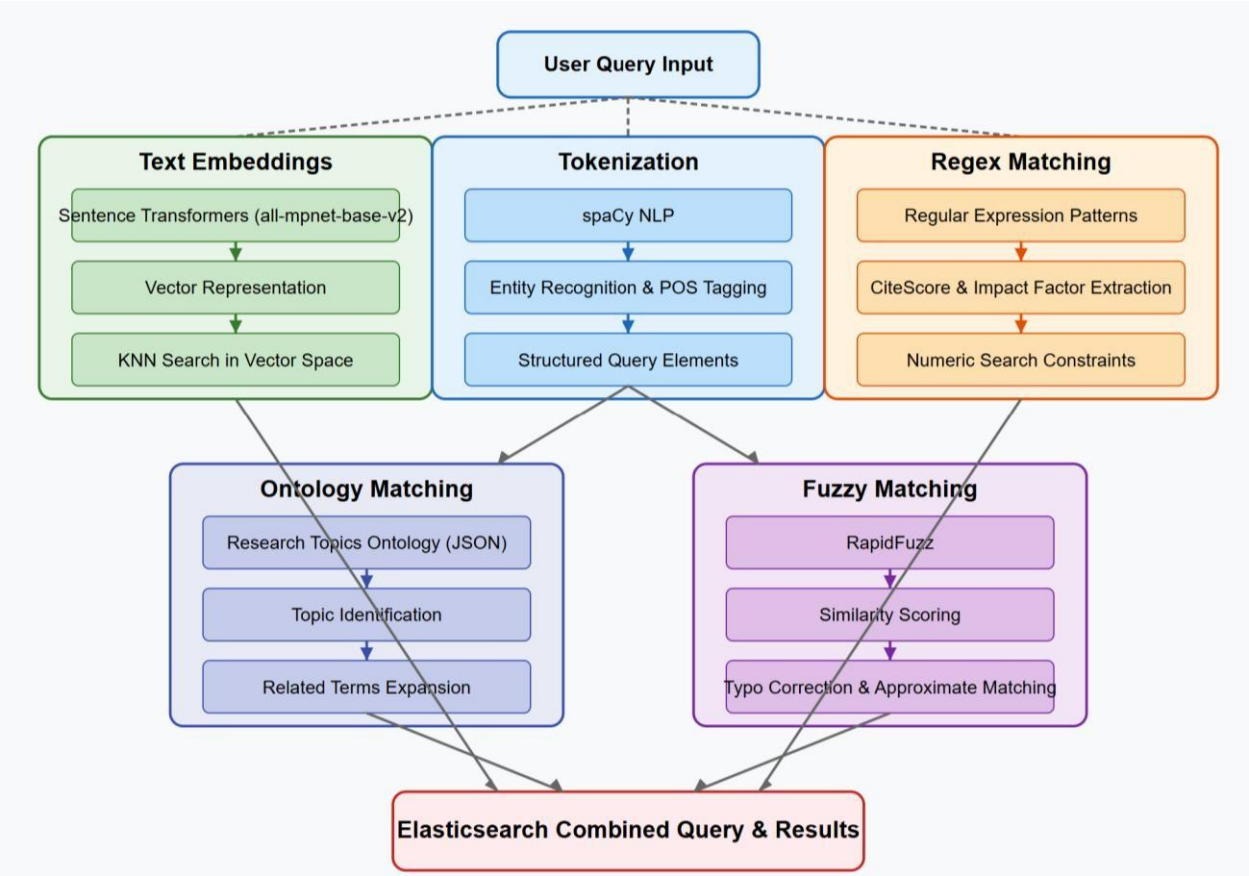
Indexing plays a crucial role in enhancing both the retrieval speed and accuracy of the proposed search system. By leveraging Elasticsearch's powerful indexing capabilities, each journal entry is transformed into a structured, searchable format that includes both semantic embeddings and token-based keywords. This dual-layer indexing—semantic and lexical—ensures that the system can quickly locate relevant records without scanning the entire dataset. The use of Sentence Transformers allows journal abstracts and titles to be embedded as dense vectors, enabling fast K-Nearest Neighbors (KNN) retrieval based on semantic similarity. Simultaneously, traditional inverted indices built from tokenized text ensure rapid and precise Boolean search matching. The combination of these approaches minimizes search latency and increases the relevance of results by aligning both user intent and exact keyword matches. Additionally, indexing numeric values such as CiteScore and Impact Factor facilitates real-time filtering and ranking, allowing users to receive not just fast but also contextually accurate results.

4.4 AI & NLP Technologies

Artificial Intelligence (AI) and Natural Language Processing (NLP) have become foundational pillars in developing intelligent information retrieval systems. These technologies enable machines to interpret, understand, and manipulate human language, making them essential for extracting meaning from large text corpora. Recent advancements in deep learning have led to the emergence of transformer-based models, such as BERT and its variants, including Sentence Transformers [28] [29], which significantly enhance semantic understanding through dense vector embeddings. In parallel, libraries like spaCy offer efficient syntactic analysis via tokenization, part-of-speech tagging, and named entity recognition, facilitating deeper linguistic analysis [30]. Moreover, ontology-based matching allows systems to align user input with domain-specific knowledge structures, improving query interpretation in specialized fields like scientific research [30]. Techniques like fuzzy string matching (e.g., RapidFuzz) provide error tolerance, handling misspellings and variations in terminology (Seifert, 2022). Additionally, regular expressions (regex) continue to play a crucial role in extracting structured data, such as citation scores or numerical indicators, from unstructured text. Together, these tools form a robust NLP toolkit for building intelligent, high-performing search systems.

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



Architecture of a Semantic and Structured Search System Using NLP and AI

The diagram illustrates the comprehensive architecture of an intelligent journal search system that integrates multiple Natural Language Processing (NLP) techniques. At the core of this pipeline is the use of Sentence Transformers (all-mpnet-base-v2) for generating semantic embeddings of user queries, enabling high-precision similarity search via K-Nearest Neighbors (KNN). To enhance query understanding, the system utilizes spaCy NLP for tokenization, part-of-speech tagging, and entity recognition—crucial for structuring the query contextually. Further semantic enrichment is achieved through ontology matching, where keywords are aligned with a pre-defined JSON ontology to accurately identify and expand research topics. Fuzzy matching, implemented via RapidFuzz, introduces tolerance for typographical errors and approximate terms, boosting retrieval robustness. Finally, regex matching is applied to extract numeric constraints such as CiteScore and Impact Factor, allowing the system to perform more granular filtering. Together, these components form a cohesive workflow that enhances both the accuracy and resilience of scholarly search results.

4.5 Search Process Workflow

The search process in modern information systems leverages a combination of semantic, syntactic, and rule-based techniques to improve retrieval accuracy. Recent advances in neural search utilize transformer models like MPNet to generate query embeddings that capture contextual meaning, enabling more precise vector-based retrieval [31]. When semantic search alone is insufficient, linguistic preprocessing tools such as Stanza [32] provide tokenization and dependency parsing to extract structured query components. For domain-specific searches, knowledge graph alignment techniques map user queries to structured ontologies, enhancing conceptual recall [33]. To improve robustness,

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

approximate string matching libraries like Jellyfish handle typographical variations, while learned pattern extraction methods (e.g., neural regex generators) outperform traditional regular expressions in complex text mining tasks [34]. By combining these approaches, modern search systems achieve state-of-the-art performance across both general and domain-specific applications.

The "Search Process Workflow" diagram illustrates a multi-phase strategy for advanced information retrieval that integrates semantic search, natural language processing, ontology-based reasoning, and Boolean query formulation. The process begins with user query input, which is converted into a semantic embedding using sentence transformers to enable a K-Nearest Neighbors (KNN) search in a vector space. If this initial semantic search returns a sufficient number of results, they are presented directly to the user. However, if the results are insufficient, the workflow proceeds through additional refinement steps. In the second stage, the query undergoes tokenization using the spaCy NLP library, allowing for the extraction of structured linguistic elements. Simultaneously, ontology matching is performed to align query terms with a predefined research topics ontology, facilitating topic identification and conceptual expansion. The third stage incorporates fuzzy matching techniques, such as RapidFuzz, to account for approximate or typographically varied inputs, and regular expressions are used to extract numeric indicators such as citation scores or impact factors. Finally, a Boolean query is constructed using the enriched components from the prior steps, and a final search is executed to retrieve the most relevant documents. The combination of semantic, lexical, and structured rule-based techniques within this workflow ensures a comprehensive and adaptive search mechanism, particularly well-suited for complex academic and scientific query scenarios.

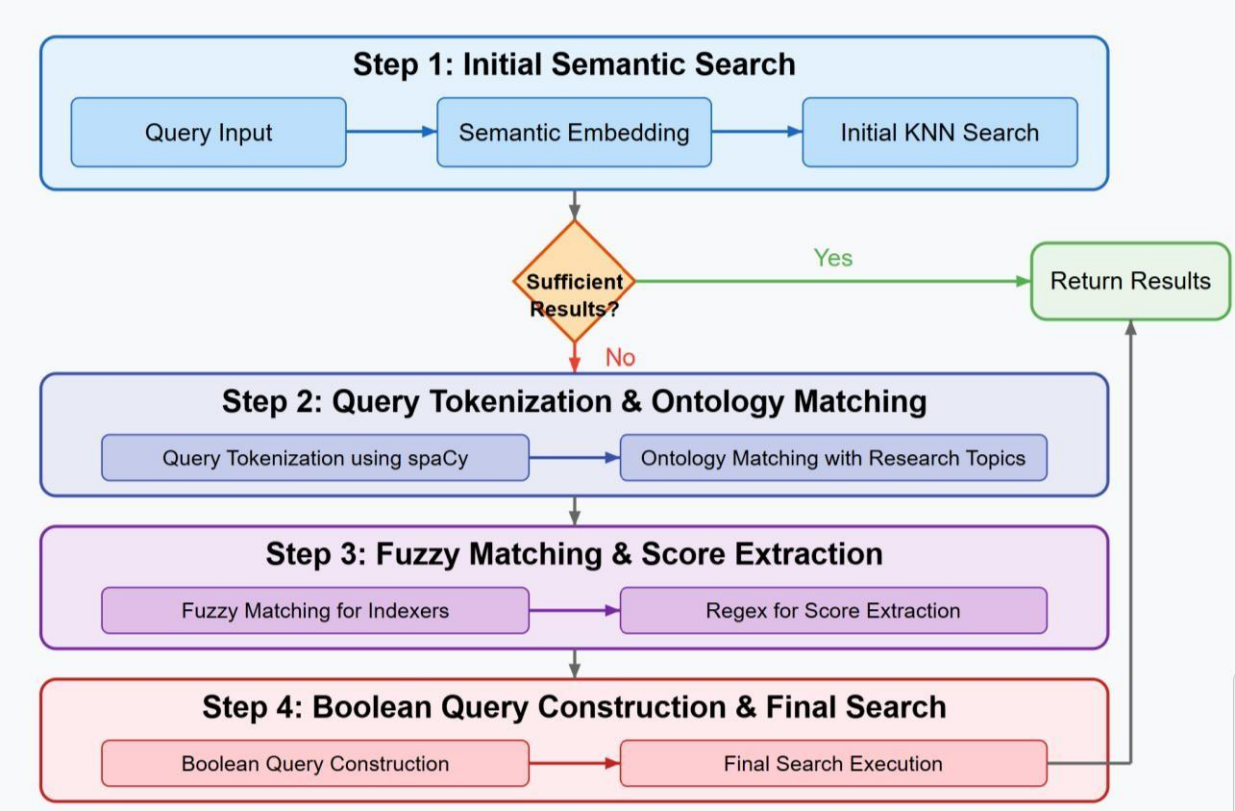


Fig: Search Process Workflow

The hybrid retrieval system processes user queries through a multi-stage pipeline that strategically combines semantic and lexical approaches. The process begins by transforming the query into a dense

vector representation using a pre-trained embedding model (e.g., BERT) and performing an initial K-nearest neighbors search in the embedding space. If results are deemed insufficient based on similarity thresholds, the system activates a fallback path involving query tokenization, ontology-based term expansion, and fuzzy matching to handle domain-specific terms and variations. The system then constructs a complex Boolean query in Elasticsearch, integrating evidence from both semantic and lexical paths through carefully weighted must, should, and filter clauses. This hybrid architecture leverages the contextual understanding of semantic embeddings while maintaining precision through ontological constraints and fuzzy term matching, addressing limitations of purely semantic or lexical approaches. The final execution against the Elasticsearch index produces relevance-ranked results that balance conceptual understanding with exact term matching, though the system requires careful tuning of semantic thresholds, ontology coverage, and query construction parameters to optimize performance. This pipeline demonstrates how modern IR systems can combine neural and symbolic techniques to handle diverse query types while mitigating the vocabulary mismatch problem.

7. CONCLUSION:

This study presents a comprehensive and scalable framework for semantic academic journal search by integrating advanced AI and NLP techniques into a hybrid retrieval system. By leveraging sentence transformer models for contextual embeddings, spaCy for linguistic preprocessing, and ontology-based expansion for domain alignment, the proposed system demonstrates an effective approach to understanding and responding to complex natural language queries. Fuzzy matching and regex-based numeric extraction further enhance the robustness of the query processing pipeline, allowing for greater tolerance to input variability and precision in filtering by citation metrics. The use of Elasticsearch enables efficient indexing and real-time retrieval, while the modular architecture ensures the system remains adaptable to future enhancements and growing data volumes.

Unlike traditional keyword-based systems, this solution bridges the semantic gap between user intent and academic content, supporting more accurate, context-aware discovery of peer-reviewed journals. The contributions of this research not only advance the field of intelligent information retrieval but also offer practical value for researchers, librarians, and academic institutions aiming to streamline access to high-impact scholarly resources. Future work may explore the integration of cross-lingual capabilities, personalized search recommendations, and adaptive learning mechanisms to further optimize performance and user experience. Ultimately, this project underscores the transformative potential of AI-driven search systems in reshaping how knowledge is accessed and utilized in academic environments.

8. REFERENCES

[1] L. Bornmann and R. Mutz, “Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references,” *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 11, pp. 2215–2222, Nov. 2015, doi: 10.1002/asi.23329.

[2] M. Gusenbauer, “Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases,” *Scientometrics*, vol. 118, no. 1, pp. 177–214, Jan. 2019, doi: 10.1007/s11192-018-2958-5.

[3] J. Beel, B. Gipp, S. Langer, and C. Breitinger, “Research-paper recommender systems: a literature survey,” *Int. J. Digit. Libr.*, vol. 17, no. 4, pp. 305–338, Nov. 2016, doi: 10.1007/s00799-015-0156-0.

[4] W. Huang, S. Kataria, C. Caragea, P. Mitra, C. L. Giles, and L. Rokach, "Recommending citations: translating papers into references," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, Maui Hawaii USA: ACM, Oct. 2012, pp. 1910–1914. doi: 10.1145/2396761.2398542.

[5] P. Mayr and A. Scharnhorst, "Scientometrics and information retrieval: weak-links revitalized," *Scientometrics*, vol. 102, no. 3, pp. 2193–2199, Mar. 2015, doi: 10.1007/s11192-014-1484-3.

[6] T. Catarci and S. Kimani, "Human-Computer Interaction View on Information Retrieval Evaluation," in *Information Retrieval Meets Information Visualization*, vol. 7757, M. Agosti, N. Ferro, P. Forner, H. Müller, and G. Santucci, Eds., in Lecture Notes in Computer Science, vol. 7757. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 48–75. doi: 10.1007/978-3-642-36415-0_3.

[7] B. J. Jansen, D. L. Booth, and A. Spink, "Determining the user intent of web search engine queries," in *Proceedings of the 16th international conference on World Wide Web*, Banff Alberta Canada: ACM, May 2007, pp. 1149–1150. doi: 10.1145/1242572.1242739.

[8] H. Schütze, C. D. Manning, and P. Raghavan, "Introduction to information retrieval cambridge university press cambridge," 2008.

[9] C. Zhai and S. Massung, *Text data management and analysis: a practical introduction to information retrieval and text mining*. Morgan & Claypool, 2016. Accessed: Apr. 22, 2025. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=WoKkDAAAQBAJ&oi=fnd&pg=PR15&dq=Zhai,+C.,+%26+Massung,+S.+(2016).+Text+Data+Management+and+Analysis:+A+Practical+Introduction+to+Information+Retrieval+and+Text+Mining.+ACM+Books.&ots=l-edHds3qq&sig=eGiB6Fmh0qbLtfVLbT98UQaUIgc

[10] S. AlGhozali and S. Mukminatun, "Natural Language Processing of Gemini Artificial Intelligence Powered Chatbot," *Balangkas Int. Multidiscip. Res. J.*, vol. 1, no. 1, pp. 41–48, 2024.

[11] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength natural language processing in python," 2020.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186. Accessed: Apr. 22, 2025. [Online]. Available: https://aclanthology.org/N19-1423/?utm_campaign=The%20Batch&utm_source=hs_email&utm_medium=email&_hsenc=p2ANqtz-_m9bbH_7ECE1h3lZ3D61TYg52rKpifVNjL4fvJ85uqggrXsWDBTB7YooFLJeNXHWqhvOyC

[13] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 27, 2019, *arXiv*: arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084.

[14] G. Antoniou and F. V. Harmelen, "A Semantic Web Primer, 2nd edn, Cooperative Information Systems." MIT Press, Cambridge, MA, 2008.

[15] C. Bizer *et al.*, "Dbpedia-a crystallization point for the web of data," *J. Web Semant.*, vol. 7, no. 3, pp. 154–165, 2009.

[16] Z. Tong, *Elasticsearch: The Definitive Guide*. O'Reilly, 2015.

[17] C. Zhai and J. Lafferty, "A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 268–276, Aug. 2017, doi: 10.1145/3130348.3130377.

[18] M. Shamsitdinova, D. Khashimova, N. Niyazova, U. Nasirova, and N. Khikmatov, "Harnessing AI for Enhanced Searching in Digital Libraries: Transforming Research Practices," *Indian J. Inf. Sources Serv.*, vol. 14, no. 3, pp. 102–109, Sep. 2024, doi: 10.51983/ijiss-2024.14.3.14.

[19] H. Xu, X. Li, J. Tupayachi, J. J. Lian, and O. A. Omitaomu, "Automating Bibliometric Analysis with Sentence Transformers and Retrieval-Augmented Generation (RAG): A Pilot Study in Semantic and Contextual Search for Customized Literature Characterization for High-Impact Urban Research," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Advances in Urban-AI*, Atlanta GA USA: ACM, Oct. 2024, pp. 43–49. doi: 10.1145/3681780.3697252.

[20] L. Zhang, C. Hu, and Z. Quan, "NLP-Powered Repository and Search Engine for Academic Papers: A Case Study on Cyber Risk Literature with CyLit," 2024, *arXiv*. doi: 10.48550/ARXIV.2409.06226.

[21] G. Samardzhiev and M. Nisheva-Pavlova, "Application of Machine Learning and Natural Language Technologies in Building Semantic Search Systems: Case Study of a Virtual Legal Assistant," in *2023 International Conference on Big Data, Knowledge and Control Systems Engineering (BdKCSE)*, Sofia, Bulgaria: IEEE, Nov. 2023, pp. 1–7. doi: 10.1109/BdKCSE59280.2023.10339730.

[22] M. Fadaee, O. Gureenkova, F. R. Barrera, C. Schnober, W. Weerkamp, and J. Zavrel, "A New Neural Search and Insights Platform for Navigating and Organizing AI Research," 2020, *arXiv*. doi: 10.48550/ARXIV.2011.00061.

[23] R. Patil, R. Jagtap, S. Yeolekar, V. Shinde, R. Sonawane, and P. Kadam, "Semantic Embedding-Based Recommender System for Research Paper Discovery and Subject Area Prediction," in *2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, Pune, India: IEEE, Aug. 2024, pp. 1–5. doi: 10.1109/ICCUBEA61740.2024.10774889.

[24] Shishir Biyyala, Sai Charan Tokachichu, and Sudheer Chennuri, "AI-Powered Search Systems : Integrating Machine Learning with Search Technology for High-Scalability Applications," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 10, no. 6, pp. 74–81, Nov. 2024, doi: 10.32628/CSEIT24106155.

[25] N. F. Ali, Md. M. Mohtasim, S. Mosharrof, and T. G. Krishna, "Automated Literature Review Using NLP Techniques and LLM-Based Retrieval-Augmented Generation," 2024, *arXiv*. doi: 10.48550/ARXIV.2411.18583.

[26] R. Anil, R. Joseph, R. Pius, S. B. K, and G. Swati Sampatrao, "Intelligent System for Research Article Recommendation: A Knowledge Graph-based Approach," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, Pune, India: IEEE, Apr. 2024, pp. 1–7. doi: 10.1109/I2CT61223.2024.10544156.

[27] T. K. Wijaya, E. D'Amico, G. Fodor, and M. V. Loureiro, "Rs4rs: Semantically Find Recent Publications from Top Recommendation System-Related Venues," in *18th ACM Conference on Recommender Systems*, Bari Italy: ACM, Oct. 2024, pp. 1174–1176. doi: 10.1145/3640457.3691696.

[28] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 27, 2019, *arXiv*: arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084.

[29] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 27, 2019, *arXiv*: arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084.

[30] R. A. Mishra, A. Kalla, A. Braeken, and M. Liyanage, "Privacy protected blockchain based architecture and implementation for sharing of students' credentials," *Inf. Process. Manag.*, vol. 58, no. 3, p. 102512, 2021.

[31] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 16857–16867, 2020.

[32] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages," Apr. 23, 2020, *arXiv*: arXiv:2003.07082. doi: 10.48550/arXiv.2003.07082.

[33] J. Pujara, H. Miao, L. Getoor, and W. Cohen, "Knowledge Graph Identification," in *Advanced Information Systems Engineering*, vol. 7908, C. Salinesi, M. C. Norrie, and Ó. Pastor, Eds., in Lecture Notes in Computer Science, vol. 7908. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 542–557. doi: 10.1007/978-3-642-41335-3_34.

[34] N. Locascio, K. Narasimhan, E. DeLeon, N. Kushman, and R. Barzilay, "Neural Generation of Regular Expressions from Natural Language with Minimal Domain Knowledge," Aug. 09, 2016, *arXiv*: arXiv:1608.03000. doi: 10.48550/arXiv.1608.03000.