

Review

Innovating Patent Retrieval: A Comprehensive Review of Techniques, Trends, and Challenges in Prior Art Searches

Amna Ali ^{1,*}, Ali Tufail ² , Liyanage Chandratilak De Silva ² and Pg Emeroylariffion Abas ^{1,*} ¹ Faculty of Integrated Technologies, Universiti Brunei Darussalam, Gadong BE1410, Brunei² School of Digital Science, Universiti Brunei Darussalam, Gadong BE1410, Brunei;
liyanage.silva@ubd.edu.bn (L.C.D.S.)

* Correspondence: aamna.hamed@gmail.com (A.A.); emeroylariffion.abas@ubd.edu.bn (P.E.A.)

Abstract: As the patent landscape continues to grow, so does the complexity of retrieving relevant “prior art”, “background art”, or “state of the art” from an expanding pool of publicly available patent data, a critical step in establishing novelty. However, retrieving this information presents significant challenges due to its volume and complexity. This systematic literature review surveys patent retrieval techniques over the past decade, focusing on ‘prior art’ and ‘novelty’ searches. Adhering to the PRISMA 2020 guidelines, our research includes 78 pertinent articles selected from a corpus of 1441, providing an in-depth overview of recent advancements, emerging trends, challenges, and future directions in the field of patent prior art retrieval. The review addresses six research questions: defining the current state of the art, evaluating the efficacy of various approaches, examining commonly used patent data collections, exploring the impact of semantic search and natural language processing (NLP) technologies, identifying frequently used components of patent documents, and discussing ongoing challenges in the domain of patent prior art search and retrieval. Our findings highlight the growing use of NLP to enhance the precision and comprehensiveness of patent searches, particularly on the Cross-Language Evaluation Forum for Intellectual Property (CLEF-IP) and the United States Patent and Trademark Office (USPTO) databases. Despite advancements, the specialized and technical nature of patent language continues to pose significant challenges in achieving high accuracy in patent retrieval.

Keywords: patent retrieval (PR); prior art search; patent search; survey; systematic literature review (SLR); patent analysis; patent mining



Citation: Ali, A.; Tufail, A.; De Silva, L.C.; Abas, P.E. Innovating Patent Retrieval: A Comprehensive Review of Techniques, Trends, and Challenges in Prior Art Searches. *Appl. Syst. Innov.* **2024**, *7*, 91. <https://doi.org/10.3390/asi7050091>

Academic Editors: Patrícia Ramos and Jose Manuel Oliveira

Received: 13 May 2024

Revised: 19 July 2024

Accepted: 23 July 2024

Published: 26 September 2024



Copyright: © 2024 by the authors. Published by MDPI on behalf of the International Institute of Knowledge Innovation and Invention. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the era of fast-evolving technology, intellectual property (IP) is crucial as it fosters innovation, drives economic growth, and protects the exclusive rights of inventors, advancing society at large [1]. Patents, central to intellectual property, uniquely grant inventors exclusive rights to use and commercialize their inventions for typically 20 years [2], encouraging research and development (R&D). This support enables companies to recoup investments and drive further innovation [3–5]. As technology develops exponentially, there is an increase in patent applications, leading to a massive accumulation of patent application documents [6]. Efficient retrieval and utilization of patent data are crucial, as patents not only serve as legal documents but also are pivotal in promoting economic development, creativity, and job creation [7].

Patent retrieval (PR) is the sub-domain of information retrieval (IR) focusing on creating strategies and approaches that efficiently and proficiently identify relevant patent documents in response to a given search query [8]. It involves a systematic search and extraction process through precise search queries on specialized patent databases and repositories while leveraging structured patent classification systems. Every stage of a patent’s journey requires the assistance of experts with domain-specific knowledge.

Therefore, an efficient patent retrieval system can aid the experts as well as enhance all the stages of the patent lifecycle, which involve some kind of patent retrieval.

Traditional manual patent retrieval methods, while reliable in some respects, are fraught with inefficiencies and bottlenecks. Manually sifting through massive patent databases is not only labor-intensive but also time-consuming, often resulting in incomplete and inconsistent results. Even well-trained and experienced patent officers find identifying related prior work to be a tedious and laborious operation [9], with a significant risk of overlooking important documents [10,11]. These limitations have prompted a growing interest from researchers and practitioners in developing more efficient and effective approaches for patent retrieval to streamline the process, enhance search accuracy, and accelerate access to relevant patent information.

The automation of the patent retrieval process began with basic Boolean and keyword-based searches but has evolved significantly over the years to complex context-based searches facilitated by deep learning and natural language processing (NLP). Early systems, reproducible but yet often ineffective, missed relevant documents or retrieved numerous irrelevant ones, necessitating labor-intensive and expertly crafted queries [12]. The development of more sophisticated techniques, such as the Vector Space Model (VSM), BM25, and Language Models (LM), brought improvements. These models evaluate the similarity between queries and documents [13] but often struggle with issues of scalability and fully comprehending the complex content of patents.

To address these deficiencies, state-of-the-art technologies including deep learning, machine learning, and natural language processing (NLP) have been introduced. NLP techniques excel at recognizing and analyzing semantic relationships within patent texts, producing search results that are highly relevant and contextually appropriate. Similarly, deep learning helps automate the analysis of sequences and feature extraction, enhancing the speed and scalability of retrieval systems.

These advanced methods significantly outperform traditional approaches in terms of efficiency, accuracy, and scalability. By better understanding the semantics of patent texts, they not only improve the accuracy of search results but also enhance the user experience by simplifying retrieval operations. Importantly, these advanced techniques can also be applied to curated data traditionally used in patent databases, further enhancing the effectiveness of search results. Indeed, many databases have started integrating both traditional curation methods and advanced retrieval technologies to leverage the strengths of each approach. This hybrid approach allows researchers and practitioners to concentrate more on analysis and innovation rather than on the tedious and laborious tasks of data preprocessing.

The integration of modern, computer-based search technologies represents a paradigm shift in patent retrieval. These methods address the limitations of traditional techniques and offer enhanced capabilities that are crucial for supporting the growing complexity and volume of patent data. This transition is pivotal for the future of patent searches, where efficiency, accuracy, and accessibility are paramount.

A detailed analysis of the state of the art in patent retrieval, particularly focusing on patent prior art retrieval, can direct future studies by pinpointing research gaps and the shortcomings of existing approaches. Before this work, no systematic literature review had comprehensively addressed these dimensions within the domain of patent retrieval. This systematic literature review (SLR) analyzes new approaches, incorporates recent developments, and synthesizes novel insights that contribute to a comprehensive understanding of the domain. As such, it fills a critical gap in the literature and provides a crucial foundation for subsequent empirical studies that may explore the applicability and effectiveness of the discussed strategies in more depth. This systematic approach not only brings clarity to the field but also sets a benchmark for future research, offering a scaffold upon which ongoing and future studies can build.

This SLR focuses on the last decade of research and development in the field of patent retrieval and patent prior art search, highlighting the transformative impact of technologies

like natural language processing, machine learning, and deep learning. These innovations have significantly enhanced the analysis and interpretation of patent data, improving the automation of classification, information extraction tasks, and the semantic understanding of patent texts. This has led to more accurate and relevant retrieval results, showcasing a leap in the effectiveness of modern retrieval systems.

Despite our focus on recent technological advances, the foundational work from earlier studies remains crucial. For instance, a 2004 study introduced a query-based associative document retrieval method that laid the groundwork for later algorithmic approaches [14]. Another study published in 2010 enhanced prior art retrieval by expanding queries with pseudo-relevance feedback [15]. These early methods demonstrated the potential of algorithmic approaches in enhancing patent search capabilities and inspired subsequent developments in the field. Moreover, the integration of machine learning into retrieval systems during the early 2000s allowed for the initial exploration of automated learning from data, setting a precedent for the more complex deep learning models that now dominate the field. These foundational techniques, though now evolved, continue to influence current research and development, providing a historical context that enriches our understanding of the trajectory of patent retrieval technologies.

The subsequent sections of the paper are organized as follows. Section 2: Patent Retrieval Tasks delves into various search tasks related to patent prior art retrieval, including a comprehensive review of related surveys, offering insights into their contributions and deficiencies. Section 3: Review Methodology outlines the systematic framework adopted in this systematic review as well as the main research questions guiding the process of uncovering key findings and trends in the domain. A detailed analysis of key findings and trends is given in Section 4: Results and Discussion, with limitations and future directions in the domain. Section 5: Conclusion and Future Direction concludes the paper by summarizing key findings from the systematic literature review.

2. Background

2.1. Innovation Lifecycle

Looking at the comprehensive patent trajectory, first, a company or inventor seeks patent protection for a novel idea. Then the application is drafted by a patent attorney who possesses expertise in both technical and legal aspects. This application consists of claims that define the envisioned extent of protection. A patent is granted if these claims demonstrate that the idea is unique and novel in comparison to prior art. Almost all these stages require preliminary prior art searches, even during the initial stages of patent drafting, which might aid in improving the patent application [16]. Therefore, patent retrieval plays an influential role at various stages of the patent lifecycle.

Below is a summary of the stages of the patent lifecycle involving prior art searches to support a range of activities.

Stage 1: Conceptualization

- At the beginning of the patent lifecycle, technology research is conducted to identify and incorporate cutting-edge technologies within a specific field to create novel solutions [8]. As a result, many prior art searches are conducted for patent landscaping.

Stage 2: Pre-Filing Stage

- Afterwards, in the pre-filing stage, innovators perform prior art searches to evaluate the originality or novelty of the invention before submitting a patent application [17].
- In the patent application drafting stage, lawyers and inventors write patent applications and refine claims by finding pertinent prior art.

Stage 3: Examination Stage

- During the patent examination stage, examiners analyze the novelty, non-obviousness, and industrial applicability of the invention by conducting targeted patent prior art retrieval searches.

Stage 4: Prosecution Stage

- In the patent prosecution stage, the examiner's objections are addressed by carrying out extended novelty searches on relevant prior art.

Stage 5: Post-Grant Stage

- In the post-grant stage, infringement and clearance inspections prior art searches are conducted by patent owners to find prospective infringers or ensure compliance.
- Technology watch and competitor analysis involve continuous tracking via numerous prior art searches to stay abreast of new developments and innovations made by rivals.
- Specific prior art searches are performed for portfolio assessments, licensing, and legal issues to make informed strategic decisions.
- Infringement and invalidity searches are performed on relevant prior art to lend credence to legal claims in court cases. When presenting evidence in court and developing legal arguments, these searches are crucial.
- Renewal and maintenance searches on prior art are essential for retaining patents by spotting relevant changes and guaranteeing that decisions made regarding the retention of patents are compliant.

Figure 1 shows different patent prior art searches that are needed at various stages of the patent lifecycle.

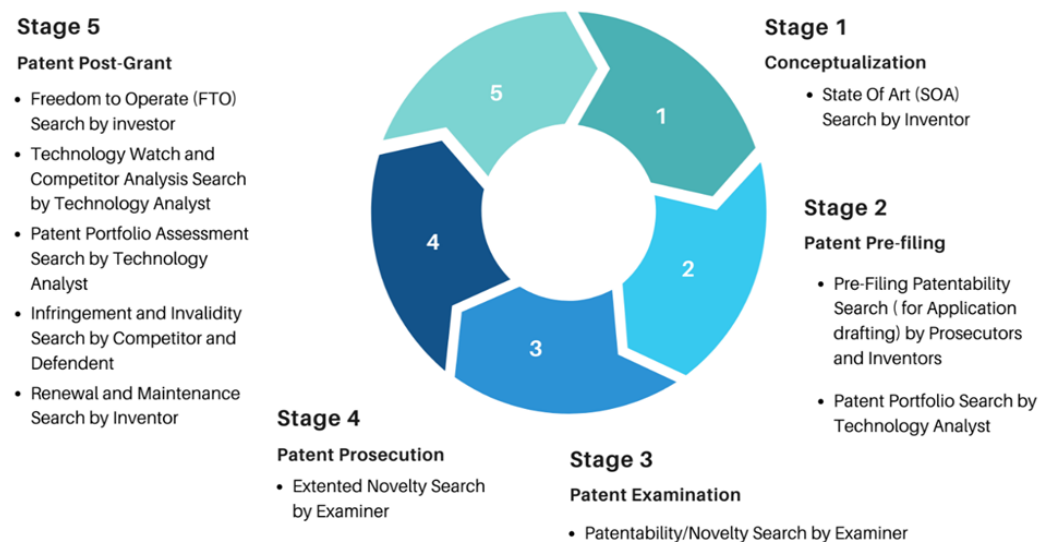


Figure 1. Patent searches to retrieve patent documents at different stages of the patent lifetime.

A high degree of domain understanding is needed for all these patent-related prior art searches and, even if such expertise is available, it needs to be coupled with extremely complex and intelligent analytics to offer users interactive and cognitive assistance [8]. The principal objectives of patent prior art search and retrieval are to guarantee compliance and shape strategic choices in legal and corporate contexts by securing, safeguarding, and leveraging patent rights. Figure 2 shows the patent searches performed at various stages of the innovation lifecycle.

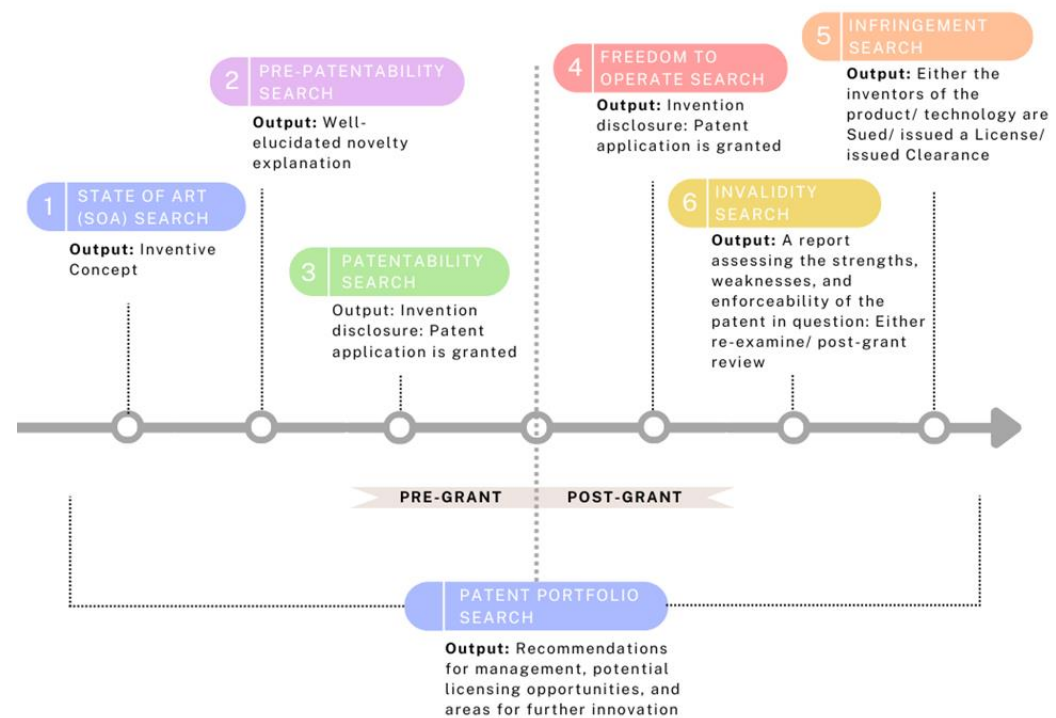


Figure 2. Innovation lifecycle.

2.2. Patent Retrieval Tasks

Patent prior art searches are specialized tasks tailored according to the characteristics of the input, such as ideas, invention disclosures, patent applications, claims, or granted patents, and the desired outputs, which may include scientific publications, a collection of patent documents, or a single document [4]. These search tasks aim to identify relevant documents to meet specific information needs and are known by various names depending on their purpose: State of the Art (SOA) search, patentability search, infringement search, freedom-to-operate search, invalidity search, and patent portfolio search. The objectives, relevance assessments, and effectiveness vary significantly based on the type of search activity.

The state of the art (SOA) search serves as a critical pre-R&D stage process, enabling inventors and researchers to gain insights into technological developments. Prior to filing a patent application, a preliminary pre-filing patentability search is normally conducted to assess novelty and non-obviousness. Subsequent to this initial decision, a comprehensive patentability search is conducted by examiners to confirm the novelty, inventive step, and industrial applicability of the invention. Before moving towards commercialization, a Freedom To Operate (FTO) search is performed to ensure that the innovation does not infringe on other active patents. Post-commercialization, competitors may engage in infringement searches to identify potential IP violations. For an in-depth exploration of these patent retrieval tasks and their contextual application, ardent readers are urged to refer to Section S2.1 Patent Retrieval Tasks in Section S2 of the Supplementary Materials.

Table 1 summarizes different aspects of various patent retrieval search tasks based on who conducts the search, the stage of the innovation lifecycle the search is taking place, the status of the patent at the time of the search, expected output from the search operation, and the literature used to perform the search on [8,18,19].

Table 1. Patent retrieval tasks.

Search Task	Conducted by	Stage	Status	Search Query Output	Literature	Part of the Document
State of the Art (SOA) Search	Inventor	Pre R&D	Pre-grant	Inventive concept	All documents available	Any
Pre-filing Patentability Search /Related Work Search	Inventor/Prosecutor	R&D/Before drafting application	Pre-grant	Well-elucidated novelty explanation/Modifications to the invention or its application	All documents available	Any
Patentability Search /Prior Art Search /Novelty Search	Prosecutor/Examiner	After drafting application	Pre-grant/Examination	Invention disclosure: Patent application is granted/Abandonment of the application if objections cannot be resolved	All patent documents are openly Accessible until the application date	Any
Freedom to Operate (FTO)	Investor	Product plan and launch: Before launching a product in the market, check commercialization risks in targeted markets	Post-grant	A product and associated procedures or technologies obtain clearance	The body of effective (active) patents in a given jurisdiction	Claims
Infringement Search	Owner/investor	Proactively done to avoid legal issues during development/In case of infringement allegations	Post-grant	Either the inventors of the product/technology are Sued/issued a License/issued Clearance	The body of effective (active) patents in a given jurisdiction	Claims
Invalidity Search	Competitor /Defendant	Before accusing of infringement (Concerns or disputes regarding the validity of a granted patent)/Performed after a patent has been granted	Post-grant	A report assessing the strengths, weaknesses, and enforceability of the patent in question: Either re-examine/post-grant review	All published patent documents accessible before the priority date of the patent in question	Any
Patent Portfolio Search	Technology analyst	Evaluate the quality, breadth, and relevance of patents within the portfolio/Technology Survey/Portfolio Survey	Pre/post-grant	Report offering insights into the overall health and strategic significance of a patent portfolio: Recommendations for management, potential licensing opportunities, and areas for further innovation	All published patent documents	Any

2.3. Existing Works on Patent Retrieval

The challenge of finding and understanding relevant patent documents amid the escalating volume of patent documents has led to the automation of patent retrieval systems, which traditionally relied on domain experts. Nevertheless, existing information retrieval (IR) approaches still struggle to effectively handle the distinctive attributes of patent documents.

Reference [12] highlights numerous research efforts at refining existing information retrieval strategies or applying standard procedures at various stages of patent retrieval. Despite the numerous efforts, patent retrieval remains an unresolved research area, with conventional information retrieval methods proving inefficient in tackling the unique challenges posed by the specific features of patent documents [12]. However, the study lacks detailed sectioning of the information gathered from the articles under review. Additionally, there are not enough graphical and tabular representations of articles based on different aspects that are crucial to patent retrieval such as which databases were utilized and which specific sections of patent documents were employed for patent retrieval. Furthermore, the study discusses patent retrieval in general and does not specifically focus on patent prior art retrieval. Another study [8] presents a thorough review of patent retrieval (PR) techniques, highlighting the need for patent domain-specific modifications in retrieval systems as traditional information retrieval approaches, such as simple web searches, are insufficient for retrieving relevant patent documents. Additionally, the review suggests a need to create interactive search tools compatible with the practices and requirements of patent domain experts [8]. Although the review attempts to categorize articles under review according to all the existing patent retrieval approaches, it does not discuss the latest state-of-the-art natural language processing (NLP) approaches. Also, it broadly mentions patent retrieval, rather than providing an in-depth review of patent prior art retrieval. Reference [20] examines the way information retrieval research has influenced and altered patent search strategies over time; however, it lacks coverage of the most recent advancements in patent retrieval techniques as it was conducted more than ten years ago [20].

In the field of patent analysis, novel approaches have been developed using deep learning and natural language processing. The use of deep learning in patent analysis was reviewed by summarizing state-of-the-art approaches and categorizing 40 research publications based on datasets and deep learning methodologies, as well as identifying possible research directions where patent analysis meets deep learning [16]. However, as the sole focus was on deep learning methodologies for patent analysis, there is a lack of comprehensive comparison with conventional or hybrid approaches. An in-depth examination of the merits and shortcomings of both deep learning and non-deep learning approaches is essential for a deeper understanding of their specific functions and effectiveness in various patent analysis tasks. Reference [21] examines the existing natural language processing (NLP) methodologies for summarizing, simplifying, and generating patent texts while acknowledging the unique challenges posed by patents in the research and development process [21]. However, the study does not directly address the patent retrieval process. Reference [22] probes the current landscape of patent analysis, its various tasks (such as forecasting technological trends, strategic technology planning, and identifying patent infringements), the prevalent tools and methodologies for efficient patent analysis, as well as the limitations of the existing tools. However, the review only briefly touches on patent retrieval tasks and predominantly centers around other aspects of patent analysis [22].

The primary objective of patent retrieval (PR) is to retrieve relevant patent documents based on a search query, which may take the form of an array of keywords, a brief, or a complete patent document. Reference [23] performs a comparative analysis of various pre-application prior art search strategies, including partial application search and query reformulation approaches from the perspective of inventors determining the patentability of their ideas before filing a full application. The complexity of lengthy queries was considered, taking into account the necessity for query reduction to remove irrelevant terms as well

as query extension to include more relevant terms. The results of this study indicated that, in terms of both writing effort and retrieval efficacy, querying with an abstract is the most balanced choice [23]. However, sole emphasis has been placed on understanding and evaluating the efficacy of various strategies linked to query reformulation, with no thorough investigation of alternative techniques used to retrieve patent papers.

In the domain of patent information retrieval, patent mining and patent retrieval are two distinct but related concepts. Patent mining uses text mining and machine learning to obtain insights from patent documents. In contrast, patent retrieval focuses on the efficient search and retrieval of relevant patent documents based on particular criteria or search queries to aid patent experts in tasks such as patent landscape analysis.

Reference [24] employs bibliometric and keyword-based network analysis to map out the evolution of patent analysis and patent mining, uncovering the key players and delineating three pivotal stages (metadata and citation analysis, cluster and network analysis, and patent mining enabled by text mining and auxiliary approaches) of development in this field, while also examining recent advancements in information retrieval and pattern analysis. A comprehensive survey of the latest trends in data mining relevant to patents, to enrich the understanding of patent analysts on the landscape of data mining, is given in reference [25]. The survey delves into multiple facets of patent mining, including patent retrieval, patent classification, and patent visualization, to offer an in-depth understanding of the area [25]. However, the survey only discusses patent retrieval as a single component of a larger investigation within the subject of patent mining. Table S1 in Section S2 of the Supplementary Materials provides detailed summaries of the shortcomings of current studies in the areas of patent retrieval [8,12,16,20–25]. This table highlights the gaps and challenges that persist in the field, underscoring the necessity for ongoing research and development in patent retrieval techniques.

This systematic literature review (SLR) on patent prior art retrieval fills a gap in existing research by systematically synthesizing and analyzing the available patent retrieval techniques from the last decade, focusing on patent “prior art” and “novelty” searches. It presents a thorough analysis of the state-of-the-art approaches in patent prior art retrieval, analyzing a substantial corpus of 78 articles. Additionally, this study exhibits clarity by thoroughly describing every database used, and its well-organized and sectioned layout makes it simple to navigate and comprehend the information presented. Also, the detailed tabular and graphical representations provide insightful illustrations of important features of patent prior art retrieval, enhancing the understanding of the results reported in the selected articles under review.

3. Review Methodology

A systematic landscape review (SLR) of patent retrieval methodologies and techniques was performed in this paper by following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [26]. Initially, specific and clear research questions were formulated to define and narrow the scope of the SLR. A suitable strategy was then developed to efficiently search common databases, utilizing various search strings to effectively retrieve relevant articles. Subsequently, articles were screened, and those most relevant were selected based on predefined inclusion and exclusion criteria for further analysis. The remaining articles after screening were then assessed based on seven pre-defined quality criteria, to appraise their quality and contributions to the technological domain.

3.1. Research Questions

A total of six research questions were formulated to define the scope and purpose of the systematic literature review (SLR). These research questions aim to guide researchers on the available and most utilized data collections, innovative and most effective state-of-the-art approaches, key performing parts of the patent documents, criteria for evaluating the quality and applicability of the approaches, the prominent challenges, and the role

of natural language processing (NLP) and semantic search in improving patent prior art retrieval techniques.

RQ1: *What innovative patent retrieval techniques have been developed and utilized for effective patent analysis?*

This question explores the evolution of patent retrieval techniques over the last decade, considering technological advancements and their impact on patent analysis. Techniques have evolved from keyword-based and Boolean searches to context-based searches to improve patent retrieval efficiency. The integration of natural language processing (NLP) and machine learning (ML) techniques with semantic analysis has enhanced search capabilities by analyzing the meaning and context of searches. Collectively, these progresses have expedited patent retrieval, providing users with more precise and accurate findings from patent archives, hence creating more intelligent, context-aware, and efficient patent retrieval systems.

RQ2: *How effective are the patent prior art retrieval techniques?*

Prior art search involves investigating publicly available patent data to determine if an invention is novel [27]. The efficacy of patent prior art retrieval systems is measured by their ability to effectively and accurately retrieve relevant patent documents in response to specific information needs or search queries. This research question reviews widely used evaluation metrics such as recall, precision, Mean Average Precision (MAP), F1-score, and accuracy to assess the performance of patent information retrieval techniques.

RQ3: *What are the most widely used patent data collections?*

In the domain of patent prior art retrieval, numerous evaluation tracks and datasets have been developed to measure the performance of systems and algorithms. This research question explores the datasets utilized by the studies surveyed.

RQ4: *What is the impact of semantic search and natural language processing on the patent retrieval process?*

Patent retrieval enhanced by semantic search and natural language processing (NLP) has transformed traditional keyword-based approaches by utilizing advanced, contextualized methods to comprehend the meaning and context behind patent search requests, allowing for the retrieval of conceptually linked patents. Such context-aware systems can scan patent data and identify relevant patent documents, including prior art and technical information, in accordance with the semantics of the request. This research question examines the impact of semantic search and NLP on enhancing the patent retrieval process.

RQ5: *Which parts of the patent document are widely used for prior art searches?*

Patent prior art search is complex due to the intricate nature of patent documents. Each component of a patent document serves a distinct function and varies in terms of significance, with claims containing the legal aspect of the inventions. Researchers utilize various components and combinations thereof, including abstracts, claims, descriptions, and references, to comprehend the context of the invention. This question explores the different parts of the patent document that have been widely employed for patent prior art searches.

RQ6: *What open challenges does prior art patent search and retrieval face?*

This question delves into the challenges within the domain of prior art search and retrieval processes. Despite the improvements in search and retrieval accuracy, numerous challenges persist, with different researchers highlighting several key challenges pertinent to the domain.

3.2. Research Strategy and Process

The methodology for article selection integrates a well-defined, unbiased search strategy crucial for ensuring the integrity and completeness of the review [28,29]. This approach combines an extensive automated search across multiple databases with a meticulous manual review of articles to identify the most relevant studies. The strategy employs composite search strings with Boolean operators ‘AND’ and ‘OR’ to optimize the retrieval process based on the defined research questions. Searches were conducted on five major academic databases: Springer Link, Google Scholar, IEEE Xplore, Science Direct, and ACM Digital Library. These searches focused on titles, abstracts, and keywords, sometimes individually or in combination, to encompass a wide range of potentially relevant articles. Table 2 provides the search strings used in the search strategy.

Table 2. Search strings used in the search strategy.

Search Strings	Objective
“patent retrieval”	Searching for articles about patent retrieval in general
“patent + retrieval AND textual”	Searching for articles about patent retrieval working on textual data
“patent + retrieval AND text”	Searching for articles about patent retrieval working on textual data, catering to the cases where articles mention just text or any other words starting with text
“prior art search”	Searching for articles mentioning only prior art search
“novelty search”	Searching for articles mentioning only novelty search
“prior + art + search AND patent + retrieval”	Searching for articles mentioning prior art search and patent retrieval
“novelty + search AND patent + retrieval”	Searching for articles mentioning novelty search and patent retrieval

Search results from the different databases were then combined before undergoing a selection process based on a thorough review of the full texts of the articles against the eligibility criteria.

3.3. Selection Requirements/Eligibility Criteria

To ensure the selection of highly relevant articles, precise inclusion and exclusion criteria were established, with meticulous attention to avoid overlooking pertinent studies. Articles retrieved using our search strings were carefully scrutinized to ensure fulfillment of the inclusion criteria; articles published before 2013 and non-English articles were excluded, while only articles in peer-reviewed conferences, workshops, and journal articles were included. Duplicate research articles were also excluded. Furthermore, full-text accessibility was a prerequisite for consideration in the final analysis. The review process focused on articles specific to text-based patent retrieval methods and techniques, excluding those related to patent ranking, hierarchy, and document structure not directly related to the retrieval of patent documents. Additionally, articles focused on non-retrieval focused patent clustering, analysis articles, review and experimental studies, works on trend analysis, technology scope, patent landscaping, and topic modeling were excluded. Table 3 provides an overview of our inclusion and exclusion criteria.

Table 3. Overview of inclusion and exclusion criteria.

Number	Criteria Name	Criteria Definition	Decision
1	Date	Years 2013–2023	Included
2	Date	Before the year 2013	Excluded
3	Language	English	Included
4	Language	Other than English	Excluded

Table 3. Cont.

Number	Criteria Name	Criteria Definition	Decision
5	Article Type	Journal and Conference/Workshop	Included
6	Article Type	Thesis, unpublished work, non-peer-reviewed, review, and experimental studies	Excluded
7	Article Access	Full access	Included
8	Article Access	Non-full access	Excluded
9	Article Relevance	Non-text-based patent retrieval and techniques not focusing on prior art search or patent retrieval	Excluded
10	Article Relevance	Text-based patent retrieval	Included

3.4. Quality Assessment

After applying inclusion and exclusion criteria to remove irrelevant articles, the remaining articles were assessed for quality using seven detailed quality criteria (QCs) addressing research robustness, publication prestige, and community value. These criteria evaluate factors such as the reputation of the publication venue, the presence of reasonable methodology, performance assessment based on real data, dataset size, acknowledgment of study limitations, citation impact, and proposed future research directions. A three-point scoring system was used for each criterion (1 for full compliance, 0.5 for partial compliance, and 0 for no compliance). The criteria and their descriptions are defined in Table 4.

Resources used to assess for Quality Criteria QC1: Publication Venue include the Scimago journal ranking [30], impact factor and indexing from Journal Citation Reports [31–33], core conference ranking [34], and Oxford List of Conferences [35]. Additionally, thorough searches were also performed on the publisher’s website for other relevant information including technical sponsors, event type, and the frequency of the event, as indications of quality. Each selected study was assessed for the quality of the proposed framework or methodology for Quality Criteria QC2. Thorough reading of the proposed framework or methodology was performed and a comparative analysis with other selected studies also aided in the quality evaluation process. Quality Criteria QC3 focuses on the presence of a proof of concept through simulation, mathematical modeling, or real-time implementations. For Quality Criteria QC4, datasets ranging from half a million to a million are deemed adequate, reflecting standards from reviewed patent retrieval research. Each study was carefully scrutinized to determine if study limitations were clearly defined for Quality Criteria QC5. For citation analysis (i.e., Quality Criteria QC6), data from GoogleScholar [36], Scopus [37], and Web of Science Citation Report [31] were considered, taking into account yearly citations and comparative article citations to gauge impact. To identify future directions for Quality Criteria QC7, each study was meticulously examined with special attention paid to the conclusion section.

Scores of each article for the seven criteria were added together to give the total score for each article. For a more accurate comparison, the final quality assessment scores were standardized using the min-max normalization technique [38]. This approach adjusts the scores to a common scale, enhancing the comparability of article quality across different metrics. The normalization formula applied converts raw scores into a standardized range, facilitating a uniform evaluation framework and ensuring that the quality assessment is both fair and reflective of each article’s relative merits within the broader research context. The following min-max normalization formula is used for the conversion:

$$D_{normalized} = \frac{D - \min(D)}{\max(D) - \min(D)} \times 100 \quad (1)$$

where $D_{normalized}$ represents the normalized score on the scale of 0–100, D represents the score to be normalized, $\min(D)$ represents the minimum score in the given set of values, and $\max(D)$ represents the maximum score in the given set of values.

Table 4. Quality criteria and scoring process.

Criterion		Description	Score	
QC1	Publication Venue	The article has been published in a reputable and high-quality venue	Full	1
			Partial	0.5
			No	0
QC2	Framework/Methodology Proposed	The article proposes a framework or methodology	Full	1
			Partial	0.5
			No	0
QC3	Proof of Concept/Performance Assessment	The article presents a proof of concept	Full	1
			Partial	0.5
			No	0
QC4	Patent Dataset Size	The article uses a dataset of reasonable size	Full	1
			Partial	0.5
			No	0
QC5	Study Limitations Defined	The article defines the limitations of their study	Full	1
			Partial	0.5
			No	0
QC6	Citations	The article has been appropriately cited	Full	1
			Partial	0.5
			No	0
QC7	Future Direction Defined	The article clearly defines the future direction	Full	1
			Partial	0.5
			No	0

4. Results and Discussion

This section presents an extensive discussion about the outcomes of the systematic analysis, structured into four key sections: “Article Selection”, “Research Questions Findings”, “Quality Assessment”, and “Research Trends and Key Findings”. In “Article Selection”, the step-by-step process used to select articles is elucidated, highlighting the academic database sources utilized, the filtering process, and the methodology employed. “Research Questions Findings” presents an analysis and findings of the selected articles in light of the defined research questions. “Quality Assessment” evaluates the quality of the included studies using established criteria and discusses both the qualitative and the quantitative analysis of the selected articles. Finally, “Research Trends and Key Findings” highlights emerging trends and significant insights drawn from the comprehensive examination and analysis of the selected studies.

4.1. Article Selection

Through the application of designated search strings across five key academic databases, a total of 1441 articles were initially identified. The first searches were conducted in July 2023. The selection process, guided by the specific inclusion and exclusion criteria and a thorough review of abstracts and full texts, narrowed this down to 78 articles for detailed analysis.

The methodology and results of this filtering process are detailed in Table 5 for each database, and the PRISMA flow diagram depicted in Figure 3 meticulously outlines the steps of article identification and screening. In the identification phase, a total of 561 articles were removed after applying the initial exclusion criteria such as date, language, and article type. Additionally, 37 duplicates were removed. In the first part of the screening phase, 631 articles were further removed after reading the title and the abstract. These removed articles focused on techniques for patent trend analysis, topic modeling, non-textual retrieval, etc. A total of 212 articles were sought for full access; however, 45 out of those did not have full access which led to their removal from the list of the shortlisted articles. After reading the full texts, a further 89 articles that did not qualify as per our outlined exclusion criteria were removed. Articles discussing patent retrieval techniques in general that could help in the patent prior art search were included in the final list of 78 shortlisted articles.

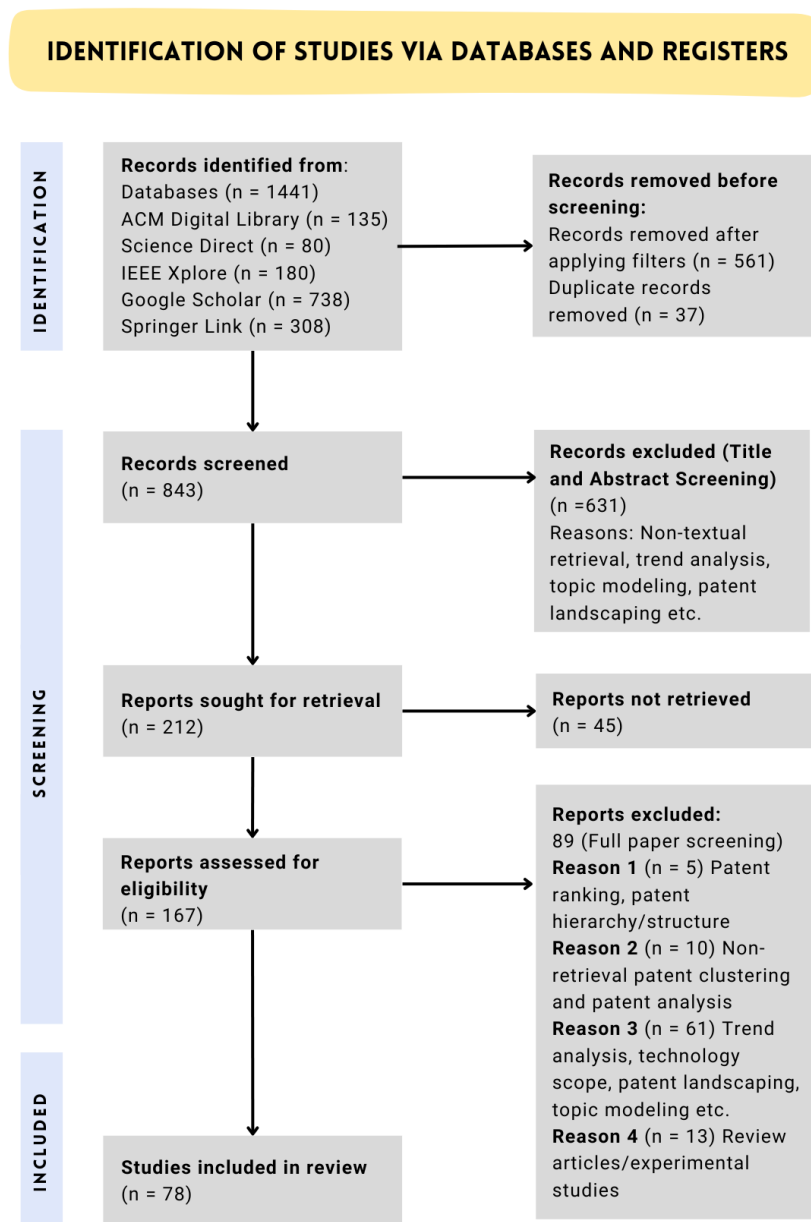


Figure 3. PRISMA flow diagram for article extraction process.

Table 5. Article retrieval per academic database.

Academic Databases	Total Articles Found	After Applying Inclusion/Exclusion Criteria and Duplicate Removal
Google	738	38
IEEE	180	14
Springer	308	16
ACM	135	9
Science Direct	80	1
Total	5	78

4.2. Research Questions' Findings

RQ1: What innovative patent retrieval techniques have been developed and utilized for effective patent analysis?

A wide range of techniques have been proposed in the literature, focusing on either general patent retrieval or retrieval specifically aiding prior art searches. These patent retrieval techniques can be divided into five (5) main categories:

1. **Bibliometric-based methods**
Bibliometric-based methods, in the context of patent retrieval, leverage patent bibliographic data and citation data to enhance the effectiveness of the patent retrieval and analysis process. Bibliographic data include the invention's title, inventor(s), assignee(s), filing date, publication date, priority date, and patent classification codes, whereas citation data comprise information on the references cited within a patent document (backward citations) and references that cite the given patent document (forward citations) [39]. Extracting citations from patent manuscripts is challenging due to the lack of a standard format for patent references [8]. For citation analysis, citation graphs and temporal patterns are commonly employed.
 - a. **Citation Graph/Network:** It is a graphical representation of relationships between patent documents by analyzing their citations. This method aids in comprehending patterns, influences, and trends within a specific area of study [39,40].
 - b. **Temporal Patterns:** These are a type of citation analysis that utilizes backward citations to show how patents cite prior works over time, providing insight into technological advancements in a particular field [41].
2. **General Information Retrieval-Based Techniques**
General information retrieval-based techniques employ principal Information Retrieval (IR) methods to search and retrieve patent documents from patent databases using methods such as keyword-based searching, Boolean queries, and relevance ranking algorithms. These include:
 - a. **Collection Selection Method:** This method helps to decide which subsection of patent documents or patent data sources should be employed during the search or analysis. The selection of pertinent data sources is based on already set parameters such as relevance to the topic, and data quality [42,43];
 - b. **Federated Search:** It enables simultaneous search through several patent databases, thereby, consolidating outcomes into one interface for swift retrieval. This method accelerates the overall information retrieval process and saves time on searches [44,45];
 - c. **Query Expansion/Construction:** It is a process of refining (adding or removing terms to fine-tune), expanding (adding terms to broaden the scope), or augmenting (adding terms to enrich coverage of a topic) the original search query to increase the inclusiveness and relevance of search outcomes [39,46–48];

- d. **Graph-Based Techniques:** These techniques leverage graph theory to represent and examine patent documents as a network of connected nodes in a graphical layout, enhancing context-aware search, retrieval, and analysis [39,48,49];
 - e. **Dynamic Ranking:** This method iteratively re-ranks search results to better reflect shifting user preferences and situations. Used by recommendation engines and search engines, dynamic ranking adaptively reorders search results to provide users with the most current and relevant content [50,51].
3. **Machine Learning-Based Techniques**
Machine learning-based techniques, in the context of patent retrieval, recognize patterns and predict document relevance to enhance patent document search, categorization, and analysis [52].
 - a. **Recommendation Based Retrieval:** This method provides tailored recommendations to users based on their search queries by utilizing methods from information retrieval, data mining, and machine learning [41,53];
 - b. **Nearest-neighbor (NN) technique:** This technique scans a set of patent documents to identify patents (nearest neighbors) that resemble a given query patent based on certain features. The similarity metric is typically determined by the distance between the feature vectors of the patent documents [54];
 - c. **Dimensionality reduction model:** This process renders a more manageable dataset for research or representation by reducing high-dimensional data to a lower-dimensional space while retaining key features (important information) [55];
 - d. **Ensemble method:** This method integrates several models to enhance the overall effectiveness of the patent retrieval system. It involves training multiple models individually and then combining their predictions to produce an improved final result [56];
 - e. **Fuzzy Logic:** This technique facilitates adaptable decision-making in situations with incomplete information by accommodating ambiguity and imprecision in patent retrieval operations [47];
 - f. **Genetic Algorithm:** This approach iteratively modifies search query parameters, including keyword combinations, and weights to better reflect the user's intent and produce more effective and comprehensive search results [57,58];
 - g. **Heuristic Meaning Comparison:** This method utilizes patterns and similarities found in patent documents to provide domain-specific insights in the form of rules to steer the search process. These approaches can be used to find synonyms, related concepts, or typical patent structures that could aid in finding more relevant patent documents [59];
 - h. **Deep Learning Techniques:** These techniques use deep learning models, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), to identify complex structures and semantic correlations within patent documents to improve overall search results [60–62];
 - i. **Clustering:** This process organizes patent documents into relevant clusters based on shared characteristics in their citation networks, metadata, or content, to aid in managing large patent databases. Additionally, it also facilitates the exploration and navigation within the network, in order to locate relevant patent documents [63–65].
4. **Natural Language Processing-Based Techniques**
 - a. **Topic Modeling:** Algorithms such as Latent Dirichlet Allocation (LDA) are used to identify generic topics or concepts in a group of patent documents by examining the density of words or terms. These algorithms categorize patent documents into topic clusters using word co-occurrence patterns, facilitating easy investigation and comprehension of key topics and trends within patent documents [61,66];
 - b. **Contextual Modelling:** These modeling techniques are employed to comprehend and generate patent textual data, analyzing technical language and domain-

specific vocabulary, and to interpret search requests. This enhances query comprehension, text summarization, and relevance scoring, ultimately improving the effectiveness of patent search and retrieval processes. Contextual modelling includes approaches such as n-gram models, Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT) [40,44,53,61,66–69];

- c. **Semantic Analysis:** This process examines the relationships between words, phrases, and sentences to comprehend and interpret the semantics (meaning) of input text. Methods used include named entity recognition, sentiment analysis, and semantic similarity [48,56,57,70–72];
 - d. **Statistical Analysis:** This technique uncovers the statistical characteristics and connections found in textual data. It supports tasks such as language modeling, information retrieval, and topic extraction by analyzing statistical traits present in the data. Included techniques are Language Models (LM) using Dirichlet smoothing and Jelinek–Mercer smoothing, Term Frequency–Inverse Document Frequency (TF–IDF), Latent Dirichlet Allocation (LDA), and n-gram models such as unigrams, bigrams, and positional information [39,46,55,58,63,73–83];
 - e. **Text Processing:** This process enables machines to examine text, understand its structure and meaning, and extract valuable information for complex tasks such as document classification and machine translation [51,62,84–87];
 - f. **Similarity Measure:** This is a comparability metric used to find similar texts or documents. It encompasses methods such as approximate nearest-neighbor techniques (identifying similar items using vector representations) and Kullback–Leibler Divergence (which measures the variation in probability distributions) [54,88];
 - g. **Tools:** Various tools like TreeTagger, OpenNLP, and Stanford CoreNLP are effective for natural language processing tasks. They offer functionalities such as tokenization, part-of-speech tagging, lemmatization, named entity recognition, sentence segmentation, and more, to conduct in-depth linguistic analysis and extract relevant information from textual data for various applications [89–91].
5. **Semantic Analysis-Based Techniques**
 - a. **Semantic Trees:** This technique organizes patent documents in hierarchical structures to illustrate semantic relationships between them as well as to facilitate systematic traversal through related patent documents [71];
 - b. **Cosine Similarity:** This measures and quantifies the degree of similarity between two patent documents' vector representations. Useful for document matching, grouping, and relevance ranking in patent retrieval systems, a high cosine similarity index suggests a higher level of similarity between documents in terms of content or attributes [90].
 6. **Other techniques**
 - a. **Claim Tree Structure:** This represents the hierarchical relationships, including comprising, linking, prepositions, and verbs, between the various components of a patent claim in a tree-like structure to provide a thorough understanding of the claim's semantics and relationships. Each component of the claim can be a single word (unigram) or a phrase (n-gram) [90];
 - b. **Patent Ontology:** This structured framework systematically arranges and represents patent-related data, such as patents, inventors, and classifications, and their relationships. It facilitates effective retrieval, analysis, and comprehension of patent data by providing a common vocabulary with shared semantics [92,93].

Table 6 illustrates the categorization of the studies according to their various techniques. It shows that 37 studies utilized general information retrieval methods, while 56 studies deployed natural language processing-based techniques. Additionally, 15 studies made use of machine learning for patent retrieval techniques, 7 studies explored bibliometric-based

techniques, and 3 studies applied other techniques. It should be noted that various studies employed more than one technique to enhance the outcome of their experiments.

Table 6. Patent retrieval techniques per category.

Techniques		Number of Studies	Study	General Category
Citation Graphs/Network		6	[39,40]	Bibliometric-Based Techniques
Patent Ranking		1	[41]	
Collection Selection method		2	[42,43]	
Federated search		2	[44,45]	General Information Retrieval Methods-Based Techniques
Query Expansion/Construction		23	[39,46–61]	
Graph-Based Techniques		8	[39,48,49]	
Dynamic Ranking		2	[50,51]	
Recommendation-based retrieval		2	[41,53]	
Nearest-neighbor (NN) technique		1	[54]	Machine Learning-Based Techniques
Dimension reduction model		1	[55]	
Ensemble method		1	[56]	
Fuzzy Logic		1	[47]	
Genetic algorithm		2	[57,58]	
Heuristic meaning comparison		1	[59]	
Deep Learning		3	[60–62]	
Clustering		3	[63–65]	
Topic Modelling		2	[61,66]	
Statistical Analysis	Bigram language Model	1	[46]	Natural Language Processing-Based Techniques
	LM (Dirichlet smoothing, and Jelinek–Mercer smoothing)	2	[78,79]	
	TF-IDF	13	[55,58,73,75–77,80,81,83,94–96]	
	Latent Dirichlet allocation (LDA)	1	[63]	
	Positional Language Model	1	[74]	
	Unigram language model	2	[39,82]	
Semantic Analysis	Doc2Vec	2	[48,70]	
	Semantic Trees	1	[71]	
	Skip-gram model	1	[57]	
	Word2vec	5	[56,72–97]	
Contextual Modelling	BERT	10	[40,44,53,66–98]	
	BiLSTM (Bidirectional Long Short-Term Memory)	1	[61]	
	Gated Recurrent Units (GRUs)	1	[69]	
	Conditional Random Fields (CRFs)	1	[68]	
Text Processing		6	[51,62,84–87]	
Similarity Measures	Approximate nearest-neighbor techniques	1	[54]	
	Kullback-Leibler Divergence	1	[88]	
Tools	Stanford CoreNLP	1	[89]	
	OpenNLP	1	[90]	
	TreeTagger	2	[91,99]	

Table 6. Cont.

Techniques	Number of Studies	Study	General Category
Semantic Trees	1	[71]	Semantic Analysis-Based Techniques
Cosine Similarity	1	[90]	
Claim Tree Structure	1	[90]	
Patent Ontology	2	[92,93]	Others

Further analysis with the help of Figure 4 shows that query expansion/construction is the most used approach in the information retrieval category. Similarly, the semantic tree-based approach and cosine similarity-based approach have both been utilized once in the semantic analysis category. Furthermore, NLP-based approaches have also been widely exploited.

RQ2: How effective are the patent prior art retrieval techniques?

When undertaking a patent retrieval task with a specific patent, whether for a new application of an innovative idea, the objective is to identify patents within a database that are pertinent to the referenced patent. This process entails a detailed examination of the patent’s claims, descriptions, and technical sphere to formulate search strategies capable of pinpointing similar inventions. The goal is to discover patents with shared technological, functional, or inventive traits, thereby providing a holistic view of the related prior art. Such an extensive search identifies not only exact matches but also patents with sufficient similarities to be deemed relevant, enriching the understanding of the patent environment surrounding the new invention. This method strives to optimize the identification of relevant patents (true positives, TP) and minimize the misidentification of irrelevant patents as relevant (false positives, FP) or the oversight of pertinent patents (false negatives, FN), ensuring the collection’s most relevant patents are precisely retrieved. Recall, precision, F1-score, Mean Average Precision (MAP), and accuracy are performance measures utilized in the literature for patent retrieval tasks. The definitions of the different parameters are given below, with Section S3.1 in the Supplementary Material providing a more thorough discussion on the topic [100–104].

Very few studies have shared comprehensive performance parameters pertinent to their proposed patent retrieval techniques. Different studies have employed different performance measures to evaluate the efficacy of their patent retrieval approaches. Some studies have focused solely on Mean Average Precision (MAP), while others have prioritized maximizing recall and enhancing precision by minimizing false positives. Additionally, several studies employ a variety of measures to provide a comprehensive evaluation of their methodologies, including precision, recall, F1 score, and others. The selection of the performance measures usually depends on the unique objectives of the research, the specific features of the dataset, and the anticipated application of the retrieval system.

Table 7 summarizes the recall, precision, MAP, F1 score, and accuracy as reported by various studies. References [50,51] reported a perfect recall rate by utilizing general information retrieval-based dynamic ranking techniques. The best precision of 99% was achieved using a nearest-neighbor-based technique [54]. The top MAP score of 92% was reported using an ML-based patent retrieval technique [54], while the highest F1 score of 95% has been obtained using an NLP-based technique [62]. Only a handful of studies have reported the accuracy of their proposed patent retrieval techniques, with the highest accuracy of 96% obtained using an ensemble-based retrieval method [56].

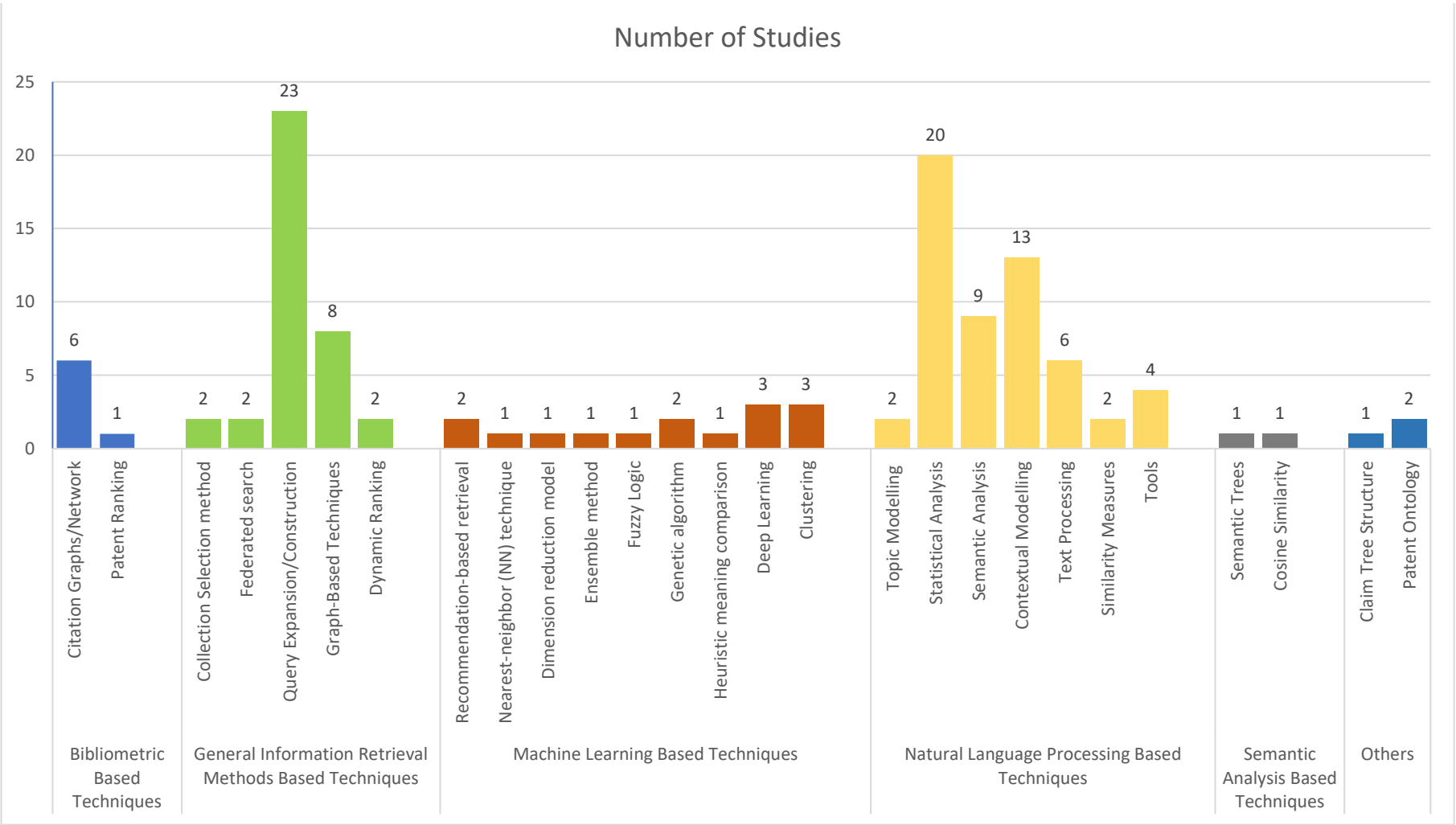


Figure 4. Patent retrieval techniques proposed by different studies.

Table 7. Performance comparison between studies.

Recall	Precision	MAP	F1 Score	Accuracy	Study
0.6595	0.554	0.105	NS	NS	[74]
NS	NS	NS	NS	NS	[70]
NS	NS	0.2727	NS	NS	[75]
NS	NS	NS	NS	NS	[47]
1	0.952	NS	NS	NS	[51]
0.2	0.65	NS	NS	NS	[68]
0.5363	0.647	0.125	NS	NS	[46]
0.611	0.58	NS	NS	NS	[105]
0.6034	0.4613	0.1731	NS	NS	[82]
0.75	NS	NS	NS	NS	[81]
NS	NS	NS	NS	0.901	[98]
0.651	0.564	0.296	NS	NS	[49]
NS	NS	NS	NS	NS	[69]
0.75	0.69	NS	NS	NS	[60]
NS	NS	NS	NS	NS	[92]
0.736	0.705	NS	0.72	NS	[93]
0.7725	0.2426	NS	0.3691	NS	[41]
NS	NS	NS	NS	NS	[43]
0.579	0.481	0.105	NS	NS	[39]
0.812	0.659	NS	0.726	NS	[67]
0.4501	0.8972	NS	0.5995	NS	[106]
0.574	NS	NS	NS	NS	[107]
0.6534	0.549	0.1312	NS	NS	[78]
0.954	0.955	NS	0.954	NS	[62]
NS	0.891	NS	NS	NS	[108]
0.944	0.0513	0.1632	NS	NS	[76]
1	0.891	NS	NS	NS	[50]
0.68	0.89	NS	0.95	NS	[61]
NS	NS	NS	NS	NS	[66]
NS	NS	NS	NS	NS	[59]
NS	0.7354	NS	NS	NS	[72]
NS	0.386	0.108	NS	NS	[109]
0.461	0.486	NS	0.473	NS	[110]
NS	NS	NS	NS	NS	[111]
NS	NS	NS	NS	NS	[112]
NS	NS	NS	0.639	NS	[53]
0.479	0.398	0.098	NS	NS	[73]
0.532	0.459	0.124	NS	NS	[85]
0.564	0.461	0.126	NS	NS	[113]
0.562	NS	0.143	NS	NS	[84]
NS	NS	NS	NS	NS	[114]

Table 7. Cont.

Recall	Precision	MAP	F1 Score	Accuracy	Study
NS	NS	NS	NS	NS	[65]
NS	NS	NS	NS	NS	[115]
0.944	0.0513	0.1632	NS	NS	[77]
0.83	NS	NS	NS	NS	[71]
NS	NS	NS	NS	NS	[40]
0.93	0.0611	0.214	NS	NS	[80]
NS	NS	NS	NS	NS	[88]
0.92753	NS	NS	0.92959	0.93166	[64]
NS	NS	NS	NS	NS	[57]
0.7136	0.65	0.1433	NS	NS	[89]
NS	NS	Na	NS	NS	[116]
0.364	0.269	0.106	NS	NS	[45]
0.9	0.88	NS	0.89	NS	[117]
0.9123	0.6682	NS	0.7714	NS	[118]
0.493	0.506	0.488	NS	NS	[58]
0.812	0.698	0.221	NS	NS	[97]
NS	NS	NS	NS	NS	[96]
NS	NS	NS	NS	NS	[44]
NS	NS	NS	NS	NS	[90]
0.83	NS	NS	NS	NS	[99]
NS	NS	NS	NS	NS	[94]
NS	NS	NS	NS	NS	[87]
0.96	NS	NS	NS	NS	[91]
NS	NS	NS	NS	NS	[42]
NS	NS	NS	NS	NS	[55]
0.605	NS	0.529	NS	NS	[119]
0.5189	0.5109	0.2585	NS	NS	[120]
0.631	0.544	0.285	NS	NS	[83]
NS	NS	NS	NS	NS	[121]
0.550	NS	0.368	NS	NS	[79]
NS	NS	0.465	NS	NS	[95]
NS	NS	NS	NS	0.965	[56]
0.49	0.99	0.92	NS	NS	[54]
NS	NS	NS	NS	NS	[86]
NS	NS	0.74	NS	NS	[63]
NS	NS	NS	NS	NS	[48]
NS	0.562	NS	NS	0.448	[122]

NS: Not Specified.

RQ3: What are the most widely used patent data collections?

Researchers must consider various factors when selecting databases for patent search and retrieval processes. Key considerations include availability, open access, and ease of use of the database. The size of the database is crucial as larger databases offer broader coverage of prior art, enabling more comprehensive searches. These extensive datasets are also vital for training machine learning and natural language processing models. Specific dataset parameters, such as publication dates, patent classification codes, and keywords, are essential for refining search queries and effectively filtering results. These parameters allow researchers to target specific time periods and technical fields effectively. Additionally, precise annotations or classifications within the database enhance the efficiency of the retrieval process. Below is a list of data collections or datasets that have been used for evaluating patent retrieval systems in the surveyed research articles:

- **CLEF-IP**

The CLEF-IP (<http://ifs.tuwien.ac.at/~clef-ip/> URL (accessed on 29 March 24)) (Conference Labs of the Evaluation Forum—Intellectual Property) is an Intellectual Property (IP) track under the umbrella of CLEF (<https://www.clef-initiative.eu/> URL (accessed on 29 March 24)) (Conference and Labs of the Evaluation Forum), which is a European series of workshops that commenced in 2001 to promote research on cross-language information retrieval (CLIR). CLEF-IP was conducted between 2009 and 2013 to evaluate the performance of patent retrieval (PR) systems. It offers datasets for various activities, including prior art searches and patent classification. The CLEF-IP data collection encompasses patent documents gathered from USPTO, EPO, and WIPO, presented in XML format with a common DTD structure. The documents include sections such as bibliographic data, abstracts, descriptions, and claims, often in multiple languages (English, German, and French), as required by the European Patent Office (EPO) for granted patents. This collection is organized in corpus and topic pools to aid the information retrieval community in comparing the efficacy of various information retrieval techniques [8,123]. A more detailed explanation of the variants of the CLEF-IP dataset is given in Table S2 in Section S3.2 of the Supplementary Material [8,123–126].

Numerous studies [39,42,43,45,46,48,49,56,66,73,74,79,82,83,85,87,89,94–118] have utilized CLEF-IP as their dataset for experimentation and verification.

- **NTCIR**

The NTCIR (<https://research.nii.ac.jp/ntcir/ntcir-12/index.html> URL (accessed on 29 March 24)) (NII Test Collection for Information Retrieval Research) workshop was first initiated in 1997 by the Japanese National Institute of Informatics to promote research in information retrieval (IR) and related fields, with a specific emphasis on cross-language information retrieval (CLIR) [8]. The Patent Retrieval Task aims to offer test sets for research on patent information processing, including retrieval and mining [127]. Table S3 in Section S3.2 of the Supplementary Material explains patent retrieval tasks in NTCIR-3, NTCIR-4, NTCIR-5, and NTCIR-6 [8,127–130].

Only one study [84] utilized the NTCIR-6 dataset.

- **TREC**

The TREC (<https://trec.nist.gov/overview.html> URL (accessed on 29 March 24)) (Text Retrieval Conference), initiated in 1992 and jointly sponsored by the National Institute of Standards and Technology (NIST) and the United States Department of Defense, is a widely recognized forum for evaluating information retrieval methods. Specifically, the TREC-CHEM track focuses on chemical patent retrieval to promote and stimulate research on chemical datasets. Table S4 in Section S3.2 of the Supplementary Material elaborates on patent retrieval-related tasks in TREC-CHEM 2009, TREC-CHEM 2010, and TREC-CHEM 2011 [8,131–133].

Two studies [58,75], reported to have utilized the TREC dataset.

- EPO

The EPO (<https://www.epo.org/en> URL (accessed on 29 March 24)) (European Patent Office), set up in 1973, is an international organization with 39 member states that is responsible for granting patents in Europe. EPO applications must be submitted in English, French, or German [134]. The EPO provides a valuable source of patent data for researchers to perform tasks related to patent retrieval and analysis, with the EPO dataset including only patents granted by the European Patent Office (EPO) since 1978 [135]. Studies such as those in references [88,117] have utilized the EPO dataset; reference [88] utilized a subset of EPO patents consisting of two million English, one million French, and one million German patents, each indexed individually for claims and descriptions, while reference [117] analyzed over a million patent applications filed with the European Patent Office (EPO) between 1982 and 2005, in conjunction with more than 20 million PubMed documents published before the beginning of 2011.

- Google Patents

Google Patents (<https://patents.google.com/> URL (accessed on 29 March 24)) provides both a search engine and public datasets, allowing users to browse and search an extensive repository of patent-related information. The Google Public Patent Database provides a huge warehouse of patent-related data and information for in-depth research and analysis and includes patents issued by 17 different patent offices, including the United States Patent and Trademark Office (USPTO) and the European Patent Office (EPO). Moreover, it features interlinked database tables to enable data-driven investigations of patent analysis and retrieval-related tasks [136]. Five studies [60,61,66,111,115] have utilized Google Patents.

- Chinese Patent

Chinese patents are managed and granted by the China National Intellectual Property Administration (CNIPA (<https://english.cnipa.gov.cn/>) URL (accessed on 29 March 24)). These patents are accompanied by datasets that provide significant information for research and analysis. CNIPA datasets include bibliographic information, publishing data, and legal data for patents, utility models, and designs from 1985 to the present. Moreover, documents can be searched using a range of search parameters, such as application number, publication number, publication date, applicant or patent holder, priority, patent agency, class, title, and abstract [137]. Only one study [121] has utilized the Chinese patents.

- IBM Almaden

The IBM Almaden (<https://research.ibm.com/labs/almaden> URL (accessed on 29 March 24)) provides a collection of more than 13 million patents stored in an Oracle Database Management System for Novartis, as part of a collaboration project between Novartis/NIBR-IT and IBM. These patents, which cover various application areas associated with life and health sciences, can be retrieved using SQL queries and are stored in an XML format. Reference [116] utilized the IBM Almaden data.

- Indian Patents

Indian patents are managed and granted by the Council of Scientific & Industrial Research (CSIR (<https://www.csir.res.in/>) URL (accessed on 29 March 24)) and the Controller General of Patents, Designs & Trade Marks (CGPDTM (<https://ipindia.gov.in/patents.htm>) URL (accessed on 29 March 24)). Patestate (<https://www.patestate.com/> URL (accessed on 29 March 24)) is an online database encompassing CSIR (Council of Scientific & Industrial Research) granted patents. The Indian patent database provides researchers with useful patent-related information, including application numbers, filing and publication dates, invention titles, international classifications, priority document details, applicant and inventor names, and abstracts [138]. Three studies [76,77,80] have utilized the Indian patents.

- MAREC

MAREC (<https://www.ifs.tuwien.ac.at/imp/marec.shtml> URL (accessed on 29 March 24)) is a large repository of over 19 million patent applications and granted patents taken from EP, WO, US, and JP databases spanning the years 1976 to June 2008. MAREC includes documents in multiple languages, including English, German, and French, with a majority being full-text documents. It facilitates research and analysis in domains that include information retrieval, natural language processing, and machine translation. Documents from numerous countries are standardized into XML format with a common citation style and patent numbering scheme. Standardized attributes include names of individuals, names of companies, dates, countries, languages, references, and detailed subject classifications. The MAREC collection consists of 19,386,697 XML files, totaling 621 GB. Two studies [44,94], have utilized the MAREC dataset.

- NIH

Primarily, the National Institutes of Health (NIH) (<https://www.nihlibrary.nih.gov/nih-subject/patents> URL (accessed on 29 March 24)) funds research and conducts studies. NIH dataset includes information (title and abstract) on extramural grants, contracts awarded, grant applications, NIH-supported organizations, NIH-funded scholars and interns via NIH programs, and biomedical manpower, covering the period from 2007 to 2010. Only one study [54], has utilized NIH dataset.

- PatentsView

The PatentsView (<https://patentsview.org/> URL (accessed on 29 March 24)) patent database is an extensive resource and an open data platform developed in partnership with the United States Patent and Trademark Office (USPTO) in 2012, focusing on data related to intellectual property (IP). This database offers features such as patent visualizations, community collaborators, an API tool, a data query builder, and bulk data download, enabling a thorough investigation and evaluation of intellectual property data. The PatentsView database offers bulk downloadable patent metadata as well as comprehensive details on granted patents, as individual files in a tab-delimited format for programmers and researchers. It has several tables containing data on applicants, assignees, attorneys, classifications, examiners, inventors, citations, and more. Only one study [68] has utilized the PatentsView dataset.

- Russian Patents

The Russian Patent (https://rospatent.gov.ru/en/products_services/search_system URL (accessed on 29 March 24)) database includes patents on various technological innovations, granted within the Russian Federation by the Russian Patent and Trademark Office (Rospatent). The Federal Service for Intellectual Property (FIPS), which is part of Rospatent, provides information on patents (encompassing inventions as well as utility models) including bibliographic data, abstracts, descriptions, claims, drawings, and legal status. The information is presented in the Russian language, but abstracts have been translated into English for broader accessibility [139]. Three studies [71,91,99] have utilized Russian Patents.

- USPTO (United States Patent and Trademark Office)

The United States Patent and Trademark Office (USPTO) (<https://www.uspto.gov/> URL (accessed on 29 March 24)) is a government office that grants patents and registers trademarks in the United States. The USPTO database enables inventors, researchers, and corporations to obtain patent-related data through search tools, including legal status information and access to full-text documents. These datasets provide comprehensive information on a range of intellectual property-related topics. Table S5 in Section S3.2 of the Supplementary Material describes the different USPTO research datasets [140].

Various studies have utilized the USPTO databases in their research [41,51,53,55,59,62,63,67,70–81,90–93,95,96,105,109–112,120].

- WIPSON

WIPSON (<https://www.wipson.com/service/mai/main.wips> URL (accessed on 29 March 24)) is managed by WIPS which is a leading intellectual property service provider in Korea. Only one study [122] has utilized WIPSON.

- TCM patents

The State Intellectual Property Office of China manages the Traditional Chinese Medicine (TCM) (<https://tcmsp-e.com/tcmsp.php> URL (accessed on 29 March 24)) patents database, established in 2001. The TCM Patent Database enables patent examiners to easily search TCM-related patents by providing access to over 19,000 bibliographic data and 40,000 TCM formulas. This database provides several search options, including quick search, advanced search, TCM formula search, and search history [141]. Only one study [64] has utilized the TCM patents,

As patentability requirements can also be influenced by non-patent data, several researchers have also utilized non-patent datasets in their studies.

- PubMed Library

The PubMed (<https://pubmed.ncbi.nlm.nih.gov/> URL (accessed on 29 March 24)) Library, administered by the National Library of Medicine, is a vast database of biomedical literature sourced from scientific journals, research articles, and books. It facilitates the search and retrieval of biomedical and life sciences literature with the intent of improving health. This database includes over 36 million citations for biomedical literature, collected from MEDLINE. MEDLINE, created by the National Library of Medicine (NLM), is a prominent bibliographic database of biomedical literature that includes citations to journal articles in the biological sciences covering medicine, nursing, dentistry, veterinary medicine, health care systems, and preclinical sciences. These citations may include links to full-text articles from the publisher and PubMed Central. Two studies [86,114] have utilized PubMed data.

- Non-Patent Datasets

Two non-patent datasets are used in the selected literature to assess the patent search and retrieval processes. The first, the Yeast (<https://archive.ics.uci.edu/dataset/110/yeast> URL (accessed on 29 March 24)) dataset, was created in 1996 and is used in biology to predict the cellular localization sites of proteins. This dataset serves classification tasks and has been employed in two studies [50,108]. The second dataset, the 20 Newsgroups (<https://www.kaggle.com/datasets/crawford/20-newsgroups> URL (accessed on 29 March 24)) dataset, comprises approximately 20,000 documents across 20 diverse newsgroups. It covers a broad range of topics from computer graphics and hardware debates to sports, politics, and religion. This dataset is particularly valuable for evaluating machine learning techniques in text-based applications such as text classification and clustering, as demonstrated in the same two studies [50,108].

Two datasets have been overwhelmingly used by about 70% of our surveyed studies. A total of 25 studies made use of the Cross-Language Evaluation Forum for Intellectual Property (CLEF-IP) dataset and another 26 studies used the dataset provided by the United States Patent and Trademark Office (USPTO), as shown in Figure 5. The remaining datasets have been used much less frequently compared to the previously reported two datasets. The Google Patents dataset has been used by five studies, whereas three studies have reported their studies using the Indian Patents dataset, and another three have utilized the Russian Patents dataset.

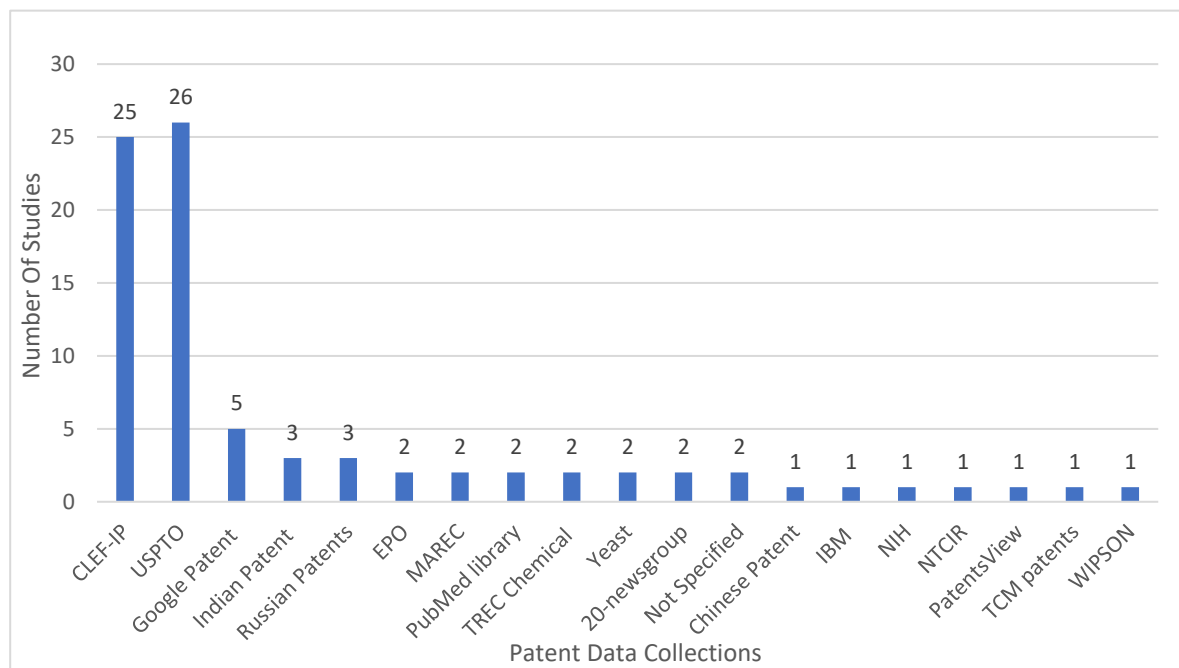


Figure 5. Dataset utilized for validation of the proposed patent retrieval techniques.

Table 8 shows more details of the CLEF-IP tracks. CLEF-IP 2011 has been widely used and has been reported by 12 studies, whereas CLEF-IP 2012 has just been used by two studies [82,87]. The table caters for the cases where a single study uses more than one CLEF-IP track. For instance, if a study uses both CLEF-IP 2009 and CLEF-IP 2010, each is reported separately in the table.

Table 8. Breakdown of CLEF-IP tracks.

Dataset	Tracks	Number of Studies
CLEF-IP	NS	1
	2009	1
	2010	9
	2011	12
	2012	2
	2013	4

The preferred databases for prior art searches in the selected studies appear to be USPTO and CLEF-IP. There could be a few possible reasons for their selection. One potential reason for the USPTO could be its comprehensive coverage of patents, incorporating a wide array of inventions. Moreover, the USPTO enjoys a remarkable reputation amongst numerous industries, making it a significant resource for research and analysis. Lastly, the USPTO provides a broad spectrum of patents and patent applications filed in the United States, making it a rich source for prior art document searches. The likely explanation for favoring CLEF-IP could be that it provides a well-curated and organized set of patents intended solely for patent prior art search tasks. Additionally, this dataset offers a well-annotated and standardized collection of patent documents, making the evaluation of prior art search and retrieval systems easier and more efficient.

RQ4: What is the impact of semantic search and natural language processing on the patent retrieval process?

Semantic search leverages natural language processing (NLP) techniques to interpret and comprehend the intent of the search query and the content of the documents being searched, facilitating the identification of more relevant and accurate results. The integration of semantic search and natural language processing (NLP) in the domain of patent retrieval enhances the efficiency and precision in locating relevant patent documents for researchers, inventors, and patent examiners. This combination enables a more refined matching process, capable of ranking complex-natured patent documents according to their semantic similarity with the search query by considering the contextual understanding of the content, which includes technical terminologies, legal jargon, and specialized acronyms.

Figure S2 in Section S3.3 of the Supplementary Material summarizes the overall NLP-based techniques reported to have been used in various surveyed research, while Table 9 shows the details of the NLP models used. For instance, SBERT is one of the most widely deployed BERT variants. Among embeddings, word embedding is the most popular embedding technique, with eight studies reported to have used it. Another significant finding from the survey is that almost all studies published from 2019 onward have incorporated NLP as their main patent retrieval technique. Moreover, the majority of those studies have included embeddings to aid the retrieval process.

Table 9. Summary of the utilized NLP models.

NLP Model	Type	Number of Studies	Embedding	Study
Kullback-Leibler Divergence		1		[88]
LM (Dirichlet smoothing, and Jelinek-Mercer smoothing)		2		[78,79]
Positional Language Model		1		[74]
Approximate nearest-neighbor techniques		1		[54]
BERT		2	Sentence and Word	[44,66]
	PLI	1	Passage	[106]
	ColBERT	1	Passage	[81]
	DistilBERT	1	Word	[53]
	ParaBERT	1	Word	[110]
	SBERT	3	Sentence	[40,67,111]
	TinyBERT	1	Word	[98]
BiLSTM		1	Word	[61]
Bigram language Model		1		[46]
Conditional Random Fields (CRFs)		1		[68]
Doc2Vec		2	Document	[48,70]
Gated Recurrent Units (GRUs)		1	Word	[69]
LDA		1		[63]
POS Tagging		6		[51,62,84–87]
Semantic Trees		1		[71]
Skip-gram model		1		[119]
TF-IDF		13		[55,58,73,75–77,80,81,83,94–96]

Table 9. Cont.

NLP Model	Type	Number of Studies	Embedding	Study
Tools	Stanford CoreNLP	1	Word Embeddings	[89]
	OpenNLP	1		[90]
	NS	1		[115]
	TreeTagger	2		[91,99]
Topic Modelling		2		[61,66]
Unigram language model		2		[39,82]
Word2vec		5	Word Embeddings	[56,72–97]

Table S6 in Section S3.3 of the Supplementary Materials offers a detailed overview of the natural language processing (NLP) techniques employed in the selected studies, providing a concise description of each method and its application within the research context [142–169]. A total of 57 studies reported to have used NLP-based techniques, and 55 studies utilized semantic or contextual searches for patent retrieval tasks. Furthermore, 21 studies employed embeddings such as word, sentence, passage, and document embeddings to retrieve the patents of interest. This information has been summarized in Figure 6.

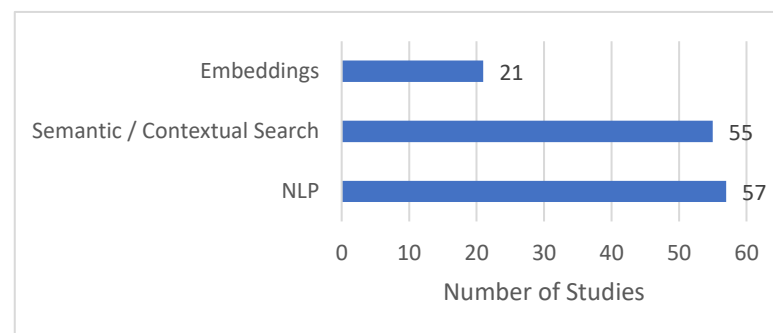


Figure 6. Utilization of semantic search and natural language processing for the patent retrieval process.

RQ5: Which part of the patent document are widely used for prior art searches?

Patents are intricate legal and technical documents, incorporating a blend of technical terminologies, legal jargon, and specialized acronyms. A patent document consists of several component sections, each serving a specific purpose and holding a different level of significance. Table S7 in Section S3.4 of the Supplementary Materials delineates the various components of a patent document, detailing their specific functions and providing explanations to enhance understanding of their roles within the overall structure of a patent.

Several studies have used complete patent documents, while others have utilized combinations of different sections, or in some cases, a single section of the patent document to retrieve patents of interest. Figure 7 summarizes the findings of the survey. A total of 14 studies have used the whole patent document to aid in the patent retrieval process, while 12 studies have utilized the title, abstract, claims, and description sections of the patent document to perform prior art and other searches. The abstract and title are also a popular combination reported to have been used by seven studies. Some studies [60,122] have made use of the title only to perform their searches. However, a total of eight studies did not specify the part of the patent document they used to perform prior art searches.

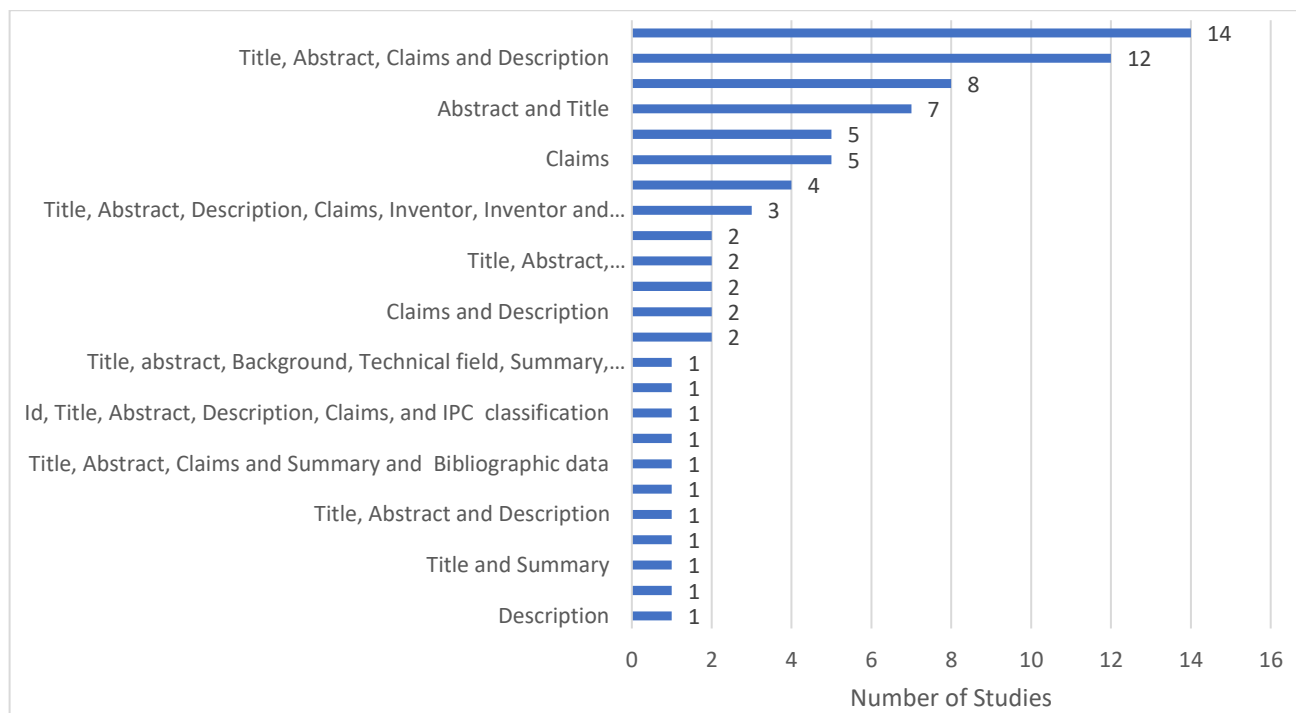


Figure 7. Summary of the part of the patent document utilized by different studies.

The preference for using full patent documents to assess the patent retrieval systems may enable a thorough grasp of the context, tackle varying levels of detail, and understand the full scope and its implications, but this requires a greater amount of time and resources. On the other hand, using specific parts of the patent documents, such as abstracts, claims, descriptions, or titles to train and evaluate a patent retrieval system is an efficient and effective way to comprehend and make use of the important information contained in the patent documents, as they offer relevance by focusing on crucial details about the invention. This method also makes system evaluation and initial relevance ranking faster. Nevertheless, selective use may overlook the contextual depth and crucial information found in full patent documents. Therefore, there should be a smart selection of specific parts of the patent documents to evaluate the retrieval systems that strike a balance between efficiency and comprehensiveness. Claims (legal scope), abstracts (concise summary), titles (short explanation of subject matter), descriptions (detailed technical information), and citations (intellectual lineage) are the sections of patent documents that can be most relevant for evaluating retrieval system performance, by considering both technical details and intellectual context.

Classification systems like the International Patent Classification (IPC) and the Cooperative Patent Classification (CPC) greatly improve patent searches by offering a standardized framework for classifying and extracting patent documents according to specific technological domains. The IPC, managed by WIPO, is used globally and involves a broad classification into sections, classes, subclasses, groups, and subgroups. It provides a systematic approach that facilitates straightforward classification and retrieval of patents, which is especially useful in traditional search environments. On the other hand, the CPC, a joint initiative by the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO), builds on the IPC with more detailed classifications, containing over 250,000 categories compared to the IPC's 70,000. This allows for even more precise and granular search outcomes, enhancing the relevance and accuracy of searches. However, while the IPC is updated every five years, the CPC system is updated monthly, reflecting its capacity to adapt more quickly to technological advancements. Nevertheless, the efficacy of these systems is contingent upon the precision of the examiners' initial classification, which can be prone to errors. Despite these challenges, both IPC and CPC codes are inval-

able resources for researchers and practitioners to perform patent searches, with the CPC offering a deeper level of detail for comprehensive investigations.

RQ6: What open challenges does prior art patent search and retrieval face?

Numerous studies have identified several key challenges in prior art searches and patent retrieval, as summarized in Table 10 from the analysis of the 78 articles. A primary concern is the ‘inconsistent document structure and formats’ [42,59,60,63,71,72,85,89,93,107], across patents. This inconsistency complicates retrieval due to the blend of structured and unstructured sections, diverse formats, and multiple languages. These variations stem from the nature of patent filings, which must accommodate a wide range of innovations and legal requirements across different jurisdictions. This leads to variations in how information is organized and presented, including differences in the use of structured (e.g., claims, abstracts) and unstructured (e.g., descriptions, drawings) sections, and the acceptance of documents in various formats and languages to suit international standards and applicant preferences. These variations pose challenges for retrieval systems in effectively interpreting the diverse content.

Table 10. Summary of the challenges that prior art searches have to face.

Challenges	Studies
Commercial patent analysis	[115]
Sophisticated patent language and vocabulary	[44,47,69,70,87,88,96,109]
Term mismatch	[44,46,61,73,78,94,113]
Information asymmetry and overload	[47,63,68,75,92,93]
Query reduction/expansion and query disambiguation	[39,61,62,74,78,107]
Inconsistent document structure and formats	[42,59,60,63,71,72,85,89,93,107]
Finding the most relevant classification codes	[117]
Large quantity of patent applications	[56,89]
Requirement of a high recall	[39,50,66,89,107]
Limitations of keyword-based search	[66,114]
Lack of data for training BERT	[44]
Retrieval results accuracy and efficiency	[57,84,111]
Examiner citations recommendation problem	[81]
Patent Ranking	[75]

Another significant obstacle is the ‘sophisticated patent language and vocabulary’ [44, 47,69,70,87,88,96,109], which arises from the need for precise descriptions and claims of innovation. The use of intricate technical, legal, and domain-specific terms aims for legal accuracy but introduces complexity in the retrieval process. Search systems must interpret and match the nuanced vocabulary used in patents, which is compounded by the vast array of patent documents across different jurisdictions with varying requirements and languages requiring translation. The challenge of accurately parsing and understanding this intricate language underscores the need for retrieval techniques sophisticated enough to effectively handle these linguistic complexities.

Several researchers have also reported ‘query reduction and query expansion’ to disambiguate the query as a significant challenge for patent retrieval [39,61,62,74,78,107]. Patent documents are inherently complex due to their lengthy nature and extensive use of technical and specialized terminology. Accurately capturing the query’s requirements and goals is crucial to retrieving the most relevant patent documents. Therefore, researchers have developed techniques to disambiguate the query, sometimes by reducing the query to remove irrelevant and noisy terms, and sometimes by expanding the query to add more relevant keywords using known external sources. Researchers must strike a balance

between oversimplification and adding complexity to the query to enhance the patent retrieval results.

‘Term mismatch’ is another crucial challenge [44,46,61,73,78,94,113], with researchers addressing the impact of semantic and conceptual variation as well as multilingual complexities on patent retrieval. Relevant documents might not be retrieved due to the use of synonyms and specialized terms in the query; therefore, researchers must develop techniques that bridge the gap between user queries and patent documents.

Some studies have reported the challenge of ‘information asymmetry and overload’ [47,63,68,75,92,93], which impacts the patent retrieval process due to imbalances in the availability of information among patent offices, applicants, and researchers, and the sheer volume of information related to patents. Applicants and researchers often lack access to cutting-edge tools and comprehensive databases, unlike patent offices, and this limited access can deter their ability to conduct exhaustive patent retrieval searches. This asymmetric access can impact the outcomes of patent searches and the overall patent application process.

The ‘requirement of a high recall’ has also been listed as one of the major challenges faced by the patent retrieval process, particularly for prior art searches [39,50,66,89,107]. High recall is necessary in prior art searches to ensure thorough scrutiny of all potential references that may impact the novelty of a new patent or invention application. Without achieving high recall, patent retrieval systems risk overlooking significant relevant patent documents and potentially facing legal issues.

Several studies have reported challenges related to the ‘limitations of keyword-based searches’ [66,114], where the focus is on overcoming the intrinsic shortcomings of keyword searches, including their inability to recognize synonyms and capture semantic meanings. Moreover, formulating complex queries involving multiple concepts or special operators, such as Boolean and proximity operators, is challenging in keyword-based searches. This can affect the outcome of patent retrieval systems. Likewise, some studies have dedicated efforts to enhancing ‘retrieval results accuracy and efficiency’. Precision and recall are both crucial in patent retrieval systems since high precision ensures that the number of retrieved patents is relevant, while high recall ensures that all relevant patents are retrieved. Nevertheless, it is a challenge to strike a balance between the two. Researchers must develop techniques that consider not just user requirements but also the context of the required search. For example, references [57,84,111] propose efficient techniques based on the Skip-gram and TF-IDF models.

Some challenges have been reported by just one study, such as ‘commercial patent analysis’ [115], typically conducted to aid strategic decision-making; ‘finding the most relevant classification codes’ [117], required as part of accurately categorizing the patent applications; ‘lack of data for training the BERT’ [44], related to the insufficiency of diverse datasets for BERT training; ‘examiner citations recommendation problem’ [81], referring to the challenges that patent examiners might face in conducting prior art searches where proper referencing is not performed on the existing patent literature; and ‘patent ranking’ [75], related to the process of assigning ranking based on relevance.

4.3. Quality Assessment

Table 11 lists the scores of each selected study against the Quality Criteria QC1 to QC7, and the final column shows the normalized score. Reference [66] is ranked 1st according to the set Quality Criteria (QCs), achieving the maximum score in 6 out of the 7 QCs. References [71,76] have accrued the lowest scores in QCs and, consequently, are the lowest in the rank.

Table 11. Quality score of studies per each quality criterion.

Study	QC1: Publication Venue	QC2: Framework	QC3: Proof of Concept	QC4: Dataset	QC5: Study Limitations	QC6: Citations	QC7: Future Directions	Total	Normalized
[66]	1	1	1	1	1	0.5	1	6.5	100
[109]	1	1	1	1	0	1	1	6	91
[111]	1	1	1	1	0	1	1	6	91
[80]	1	1	1	0.5	1	0.5	1	6	91
[46]	1	1	1	1	0	0.5	1	5.5	82
[39]	1	1	1	1	0	0.5	1	5.5	82
[62]	1	1	1	0.5	1	0.5	0.5	5.5	82
[53]	1	0.5	0.5	0.5	1	1	1	5.5	82
[121]	1	1	1	1	0	0.5	1	5.5	82
[74]	1	1	1	1	0	0.5	0.5	5	73
[49]	1	1	1	0.5	0	1	0.5	5	73
[106]	1	0.5	0.5	1	0	1	1	5	73
[78]	1	1	1	1	0	0	1	5	73
[117]	0.5	1	1	1	0	0.5	1	5	73
[70]	0.5	1	1	1	0	0	1	4.5	64
[68]	1	1	1	0.5	0	0	1	4.5	64
[82]	0.5	1	1	1	0	0	1	4.5	64
[81]	1	1	1	1	0	0.5	0	4.5	64
[69]	0	1	1	1	0	0.5	1	4.5	64
[110]	1	1	1	0.5	0	0	1	4.5	64
[73]	0.5	1	0.5	0.5	1	0	1	4.5	64
[85]	0.5	1	1	1	0	0	1	4.5	64
[113]	0.5	1	1	1	0	0	1	4.5	64
[84]	0.5	1	1	1	0	0	1	4.5	64
[114]	0.5	1	1	0.5	0	0.5	1	4.5	64
[77]	0.5	1	0.5	1	0	0.5	1	4.5	64
[116]	0.5	1	1	1	0	0	1	4.5	64
[97]	0.5	1	1	1	0	0	1	4.5	64
[96]	0	0.5	0.5	0.5	1	1	1	4.5	64
[120]	1	1	1	1	0	0.5	0	4.5	64
[95]	1	1	1	1	0	0.5	0	4.5	64
[54]	0	1	1	0.5	0	1	1	4.5	64
[75]	0.5	1	1	0.5	0	0	1	4	55
[67]	0.5	1	1	1	0	0.5	0	4	55
[72]	0.5	0.5	0.5	0.5	0.5	1	0.5	4	55
[112]	0.5	1	1	1	0	0.5	0	4	55
[45]	1	0.5	1	1	0	0	0.5	4	55
[118]	0.5	0.5	1	1	0	1	0	4	55
[58]	0.5	0.5	0.5	1	0	1	0.5	4	55
[90]	0.5	1	1	0.5	0	0	1	4	55
[99]	0	0.5	1	1	1	0	0.5	4	55
[42]	0	1	1	1	0	1	0	4	55
[55]	0	1	1	1	0	0	1	4	55
[79]	0.5	0.5	0.5	0.5	1	0	1	4	55
[63]	0.5	0.5	0.5	0.5	0	1	1	4	55
[48]	0	1	1	1	0	0	1	4	55
[122]	0	1	1	1	0	0	1	4	55
[51]	0	1	1	0.5	0	0	1	3.5	46
[92]	0.5	1	0.5	0.5	0	0	1	3.5	46

Table 11. Cont.

Study	QC1: Publication Venue	QC2: Framework	QC3: Proof of Concept	QC4: Dataset	QC5: Study Limitations	QC6: Citations	QC7: Future Directions	Total	Normalized
[41]	1	1	0.5	1	0	0	0	3.5	46
[40]	0	1	1	1	0	0	0.5	3.5	46
[64]	0	0.5	1	1	0	0	1	3.5	46
[94]	0	1	1	1	0	0	0.5	3.5	46
[59]	0	0.5	0.5	0.5	0.5	0.5	0.5	3	37
[43]	0	0.5	0.5	1	0	0	1	3	37
[107]	0	0.5	1	1	0	0	0.5	3	37
[108]	0.5	0.5	0.5	0.5	0	0.5	0.5	3	37
[50]	0.5	0.5	0.5	0.5	0	0.5	0.5	3	37
[87]	0	0.5	0.5	1	0	0	1	3	37
[56]	0	1	0.5	0.5	0	0	1	3	37
[61]	0	1	0.5	0.5	0	0.5	0	2.5	28
[47]	0.5	0.5	0.5	0	0	0	1	2.5	28
[105]	0	1	1	0.5	0	0	0	2.5	28
[60]	0	1	1	0.5	0	0	0	2.5	28
[93]	0	1	0.5	0.5	0	0	0.5	2.5	28
[65]	0	0.5	0.5	0.5	0	0	1	2.5	28
[115]	0	0.5	0.5	0.5	0	0	1	2.5	28
[91]	0.5	0.5	0.5	1	0	0	0	2.5	28
[86]	0	1	1	0.5	0	0	0	2.5	28
[98]	0	1	0.5	0.5	0	0	0	2	19
[88]	0.5	0.5	0.5	0.5	0	0	0	2	19
[89]	1	0.5	0.5	0	0	0	0	2	19
[44]	0.5	0.5	0.5	0.5	0	0	0	2	19
[83]	0	0.5	0.5	0.5	0	0	0.5	2	19
[57]	0	0.5	0.5	0.5	0	0	0	1.5	10
[119]	0	0.5	0.5	0.5	0	0	0	1.5	10
[76]	0	0.5	0.5	0	0	0	0	1	0
[71]	0	0.5	0.5	0	0	0	0	1	0

Further analysis of the top five studies [46,66,80,109,111] reveals that, except for one study [80]—which only partially fulfilled QC4 due to the use of a smaller dataset—all the other studies fully met QC1 to QC4 and QC7. These studies have been published in reputable venues, have presented and tested their frameworks on well-known and large datasets, and have highlighted future research directions. Whereas two studies [66,80], fully met QC5 by discussing the limitations of their study, none of the other studies stated any limitations. Similarly, studies [109,111] fully met QC6 with good citation rates per year; however, the rest of the studies only partially fulfilled QC6.

Upon analyzing the bottom three studies [71,76,119], it can be observed that none of the studies fully met any of the given QCs. Only QC2 and QC3 have been partially met by all the studies, as these studies either propose an incomplete framework or have not detailed the proof of concept within the study. QC4 was only partially met by reference [119] by utilizing a basic dataset; neither of the other two studies mentions datasets. All of these studies were also not published in a reputable venue, received few or no citations per year, and did not discuss any future research directions, hence not meeting QC1, QC6, and QC7.

4.4. Research Trends and Key Findings

To enhance the discussion on recent research trends and key findings in patent retrieval for prior art and novelty searches, it is crucial to align these insights with the initially set research questions. This approach ensures a structured analysis and understanding of the

evolving landscape in patent information retrieval. By examining the advancements and methodologies developed in recent studies, one can identify the direction in which the field is moving and the potential areas for future investigation. This includes recognizing the adoption of sophisticated machine learning and natural language processing techniques, addressing the challenges posed by complex patent language and diverse document formats, and the increasing need for systems capable of navigating the intricacies of global patent databases.

4.4.1. Publication Trends

Figure 8 reveals a fluctuating trend in patent retrieval research publications from 2013 to 2023, with the highest number of annual publications reaching 14 in 2015. This period of peak activity between 2013 and 2015 is, however, followed by a notable decrease in 2016 and 2017. The number of publications saw a resurgence in 2018 and 2019, only to decline again in 2020. Despite this inconsistency, a slight increase in 2021 and 2022 indicates a renewed interest, particularly in leveraging NLP techniques. As the search was initially conducted in July 2023, the publication data for the year 2023 constitutes only partial data, as it takes up to 6 months for a publication to be properly indexed in databases. Overall, this observation suggests a growing interest among researchers in exploring advanced NLP applications in patent retrieval, hinting at a potential rise in publications in the near future, as further discussed in the subsequent section.

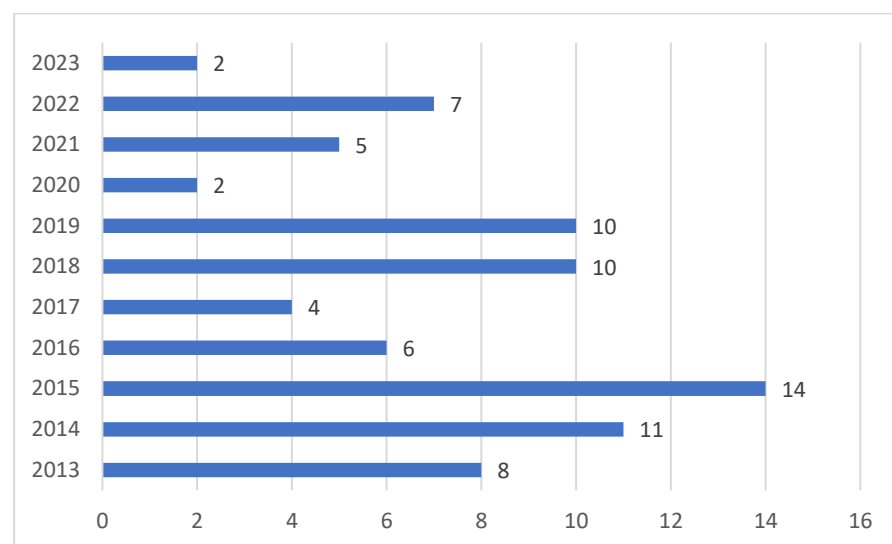


Figure 8. Number of published papers per year.

Figure 9 illustrates that a significant majority of literature in this field, approximately 66% of the overall articles, is published in conference proceedings, with the remaining 34% appearing in journals. This distribution suggests a preference among researchers for conference presentations due to the immediate feedback, networking, and visibility they offer. Conferences also facilitate multi-disciplinary interactions, providing a dynamic platform for researchers to share and refine their ideas, a factor that could explain the trend toward conference publications in the realm of patent retrieval research.

Table 12 shows that out of the top 10 publications that received the highest quality score, seven (7) are journal publications while only three are conference publications. This result reiterates the already known fact that the quality of journal papers is generally considered superior to that of conference papers. Moreover, these studies are associated with various well-known publishers including Elsevier, IEEE, Taylor & Francis, ACM, Springer, and Emerald. Similarly, Table 13 shows that out of the top 10 most cited papers, five are journal publications, with another five being conference papers. This observation indicates that citations on papers are not merely dependent on the type of publication,

despite the perceived quality of journals, but rather on the ideas presented in the publication. Additionally, Springer has the highest number of papers, four, in the top ten list, whereas IEEE, ACM, and Elsevier have three, two, and one paper, respectively.

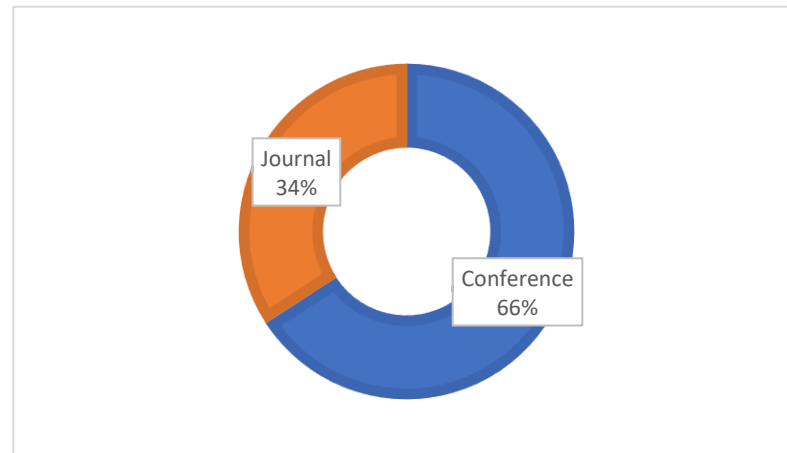


Figure 9. Summary of the types of publications.

Table 12. Top ten publications with the highest quality scores.

Ref	Publisher	Conference/Journal	Quality Score
[71]	Elsevier	Journal	100
[104]	IEEE	Journal	91
[43]	Taylor & Francis	Journal	91
[63]	IEEE	Conference	91
[46]	ACM	Conference	82
[107]	Springer	Journal	82
[77]	Emerald	Journal	82
[59]	Elsevier	Journal	82
[42]	ACM	Conference	82
[78]	ACM	Conference	73

Table 13. Top ten most cited papers.

Ref	Publisher	Conference/Journal	Citations	Normalized
[86]	IEEE	Conference	71	55
[104]	IEEE	Journal	55	91
[114]	Springer	Journal	53	55
[57]	Springer	Journal	40	55
[94]	Springer	Conference	39	55
[78]	ACM	Conference	37	73
[59]	Elsevier	Journal	33	82
[54]	ACM	Conference	31	64
[107]	Springer	Journal	30	82
[76]	IEEE	Conference	24	64

Figure 10 shows the distribution of the number of publications per publisher. IEEE accounts for a significant portion, with 31% of publications, while Springer follows closely with 28%. Emerald and Tylor & Francis, on the other hand, received a much smaller share of publications, with just 3% and 2% of the total, respectively.

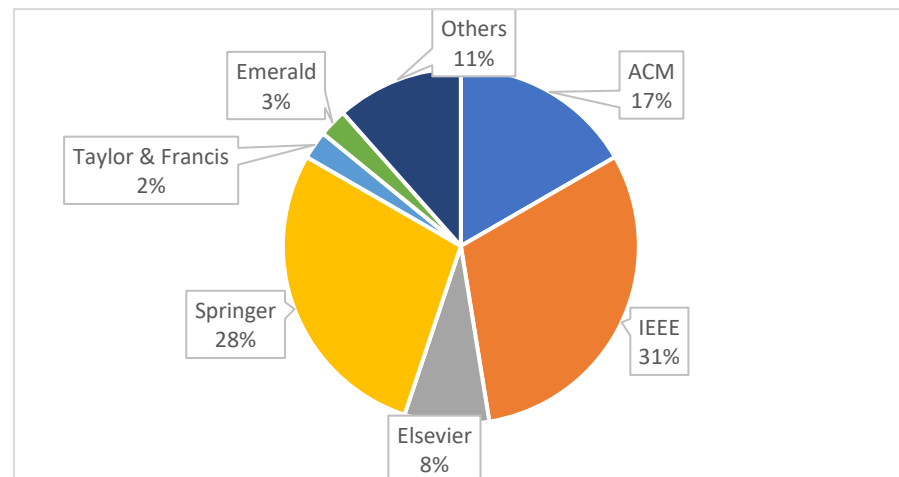


Figure 10. Number of publications per publisher.

Figure 11 summarizes the trend in the number of journal and conference publications per publisher. It is evident that IEEE, ACM, and Springer are favored venues for conference publications, likely due to their specialization in fields related to information retrieval, data mining, and artificial intelligence. This specialization is a key factor contributing to their high number of conference papers. In contrast, Elsevier and Taylor & Francis are noted for only journal publications, focusing on a wide array of disciplines, which explains their prominence in journal outputs.

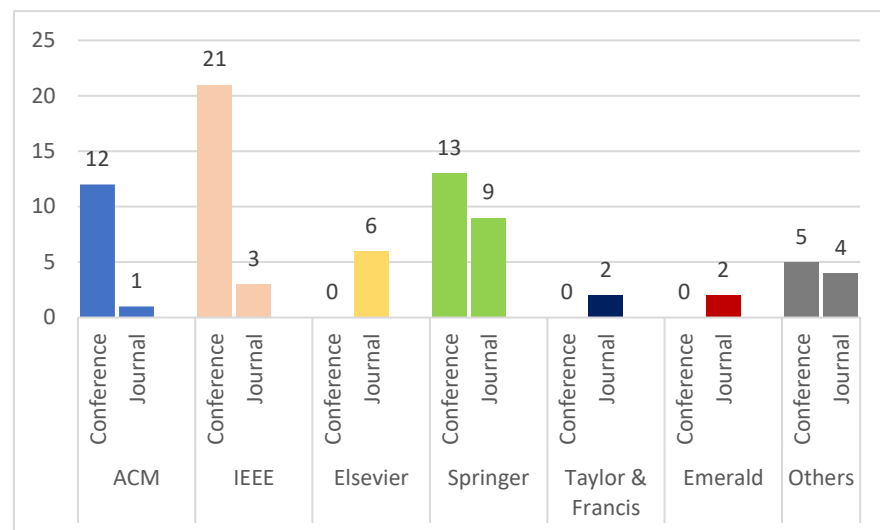


Figure 11. The number of conference and journal publications per publisher.

4.4.2. Research Questions: Key Findings

- **RQ1: Emerging Patent Retrieval Techniques**

The survey revealed that “Query Construction/Expansion” has traditionally been the most popular technique amongst researchers [60–74], assisting in patent retrieval tasks. The main focus of these techniques was to enhance recall by reducing query mismatch and ambiguity. Recently, with advancements in NLP, more sophisticated context and semantic-aware algorithms have been developed, yielding better results for patent retrieval tasks with improved precision, as seen in references [62,99,111]. Additionally, “Graph-based techniques” [40,49,108] have also been widely explored, due to their ability to represent both structured and unstructured patent data and their complex relationships, facilitating patent ranking and analysis by identifying the influential patents and refining the relevance and

quality of search results. Based on the surveyed studies, Figure 12 presents the classification of the patent retrieval and patent prior art search-based techniques. These techniques have been classified into five general categories: Bibliometric-Based Techniques, General Information Retrieval-Based Techniques, Machine Learning-Based Techniques, Natural Language Processing-Based Techniques, Semantic Analysis-Based Techniques, and Others. The sub-categories are also shown in the figure.

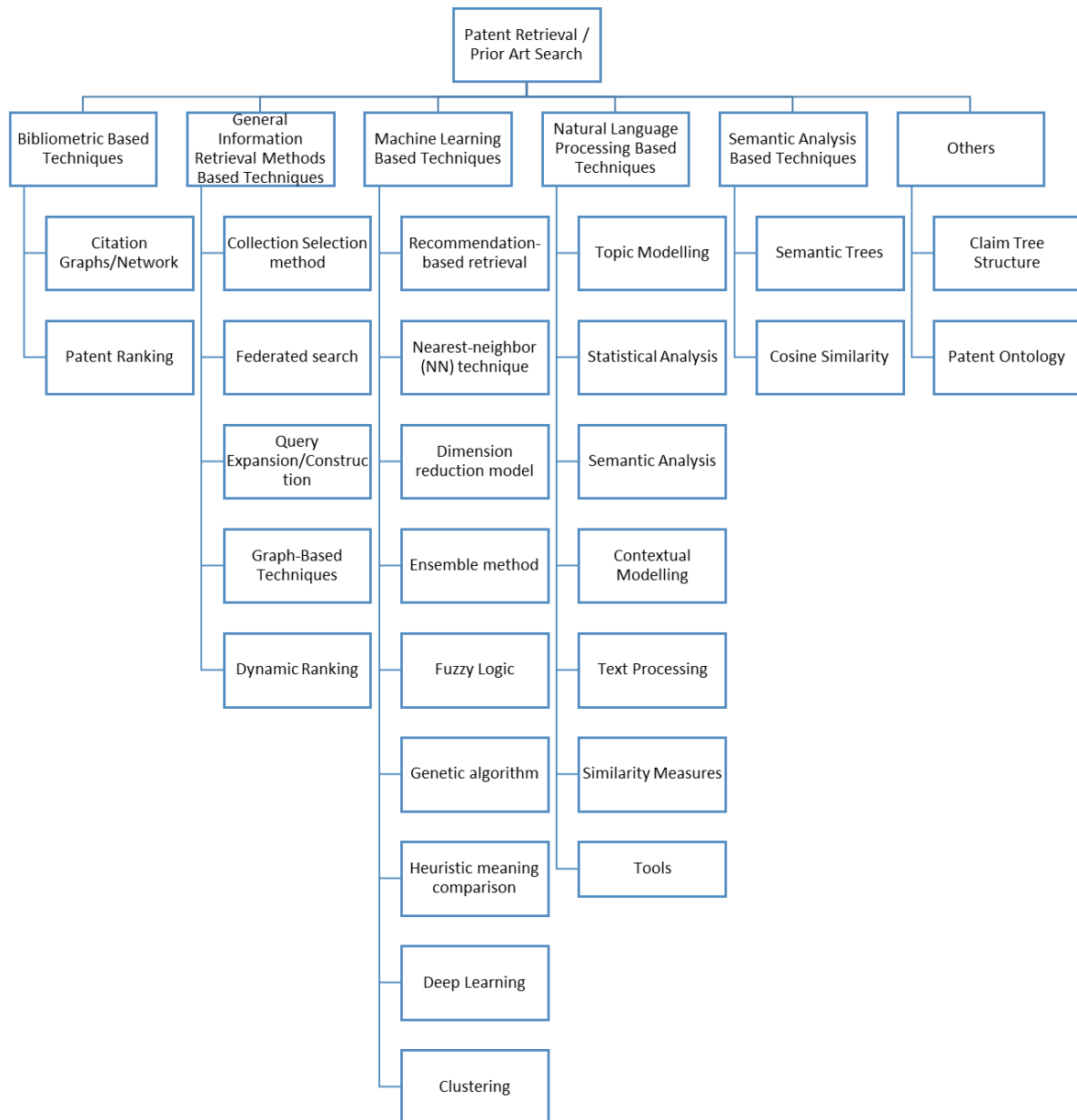


Figure 12. Classification of patent retrieval/prior art search techniques.

- **RQ2: Top Performing Patent Prior Art Retrieval Techniques**

Two studies, [50,51], based on dynamic ranking techniques, reported a perfect recall rate for retrieval, with reference [50] utilizing the Yeast dataset while reference [51] utilized the USPTO dataset to perform experimentation. This highlights the importance of dynamic ranking techniques in retrieving a high number of relevant patent documents on versatile datasets. However, precision rates of 89% and 95% were reported by these studies, respectively. The top precision of 99% was achieved using a Nearest Neighbor-based technique [54]. The use of similarity measurement to retrieve relevant patent documents

can be one of the contributing factors that yielded such high precision for Nearest Neighbor-based techniques. Balancing between precision and recall, the highest performing model employed an NLP-based technique [62], achieving an F1 score of 95%. This underscores the effectiveness of NLP techniques not only in accurately identifying relevant patent documents but also in minimizing the detection of irrelevant ones.

- RQ3: Popular Datasets

Figure 13 illustrates the use of patent datasets over the surveyed years, showcasing a strong preference for the USPTO and CLEF_IP databases among researchers. More than 70% of our surveyed studies make use of the two most popular datasets, with 26 studies making use of the United States Patent and Trademark Office (USPTO) dataset [81,92,98] and 25 studies employing the Cross-Language Evaluation Forum for Intellectual Property (CLEF-IP) [39,43,82]. These choices highlight the versatility and relevance of the corpora, as both remain favored for their comprehensive and high-quality data.

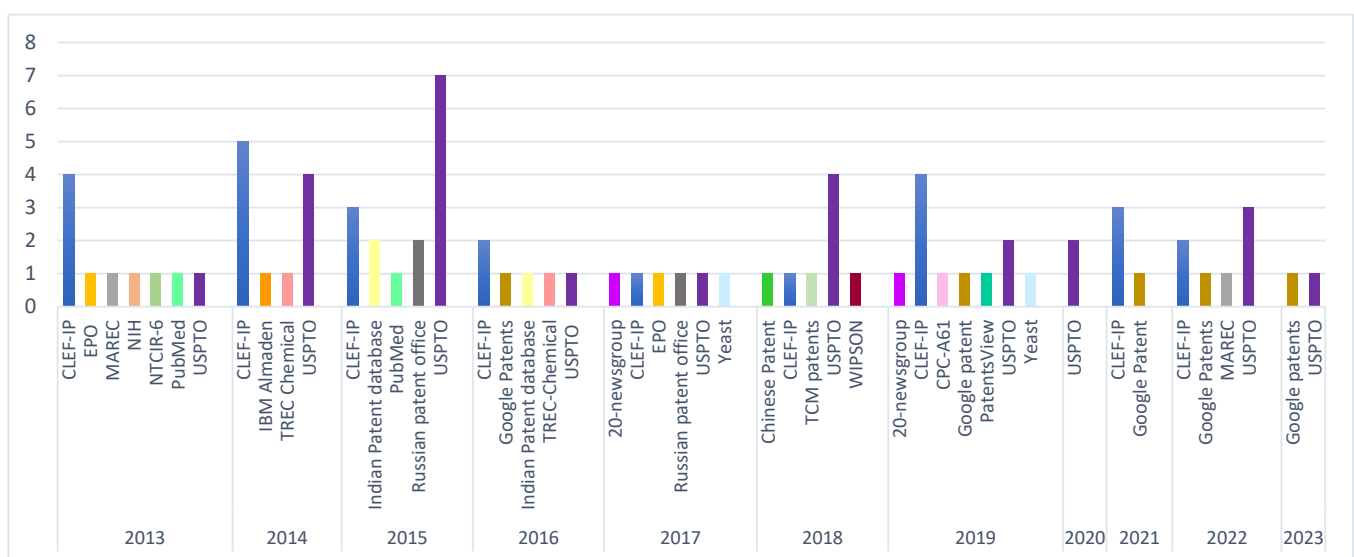


Figure 13. Patent datasets used over the years.

The USPTO database, renowned for its extensive coverage of U.S. patents and patent applications, includes a record number of up to 10 million records utilized in the most extensive study [83]. This database's breadth makes it an invaluable resource for patent prior art searches, widely recognized across numerous industries for its utility in research and analysis. Meanwhile, CLEF-IP, particularly the 2011 variant which is the most widely used, sees up to 3.5 million records employed from its datasets for a single study [69]. Designed specifically for patent prior art search tasks, CLEF-IP offers a well-curated and organized collection that provides standardized and well-annotated patent documents. This makes it easier and more efficient to evaluate prior art search and retrieval systems.

Recently, Google Patents has also begun to gain recognition from the research community due to its easy access, comprehensive coverage, and availability of structured data, signaling a potential shift in preferred resources or a broadening of the tools available for patent research. Despite the introduction of newer platforms like Google Patents, USPTO and CLEF_IP remain the cornerstones for most patent retrieval research, consistently chosen for their robust datasets that support the development of innovative prior art search and patent retrieval techniques. Their continued popularity underscores their critical role in shaping the future directions of patent retrieval research.

- RQ4: Impact of Natural Language Processing on Patent Retrieval

Many of the recent studies have shifted their focus to using NLP-based techniques for prior art searches. Our survey revealed that almost all studies published from 2019 onward

have included NLP as their main patent retrieval technique. This shift from traditional approaches to NLP may be attributed to various advantages that NLP-based techniques offer, including semantic and contextual understanding, multi-lingual support, and the availability of advanced deep learning-based algorithms. The majority of the surveyed studies have made use of embeddings to aid the retrieval process. Embeddings are vector representations of words, sentences, and documents that help the retrieval process with better semantic and syntactic understanding of data. Among the different NLP techniques, BERT is an emerging choice by researchers, deployed in a total of 10 studies. BERT provides various advantages, including pre-trained contextualized word embeddings, assisting transfer learning, and supporting deep bidirectional representation learning.

Another important trend observed in our surveyed articles is that the highest recall rates have been achieved by all those studies that deployed NLP as their main technique. For example, studies such as those in references [51,62,64,77,80,91] showed above 90% recall rates, indicating a significant impact of NLP-based approaches on patent retrieval tasks, especially for novelty and prior art searches.

Figure 14 presents a classification of NLP-based techniques. These techniques have been placed into seven categories: Topic Modelling Techniques, Statistical Analysis Techniques, Semantic Analysis Techniques, Contextual Modelling Techniques, Text Processing Techniques, Similarity Measures Techniques, and Tools. The sub-categories have also been shown in the figure.

- RQ5: Popular Patent Document Sections

Analysis of the selected studies revealed that 14 studies utilized the full patent document for prior art and other patent retrieval searches. While the full document provides comprehensive information on the patent, thereby potentially enhancing retrieval outcomes, it should be noted that utilizing the full document is resource-intensive. This is because a single full document contains a large amount of information, requiring significant computational power and storage capacity to process effectively. 13 out of those 14 studies utilized NLP as their search technique.

Another 12 studies reported utilizing “Title, Abstract, Claims, and Description” to perform patent retrieval tasks, proving the effectiveness and usability of this combination. Similar to the above, “Title, Abstract, Claims and Description” is also a common combination being utilized by NLP-based approaches [113,122].

A significant number of studies achieving a recall rate of 90% or higher utilized either the “Title” section alone or the combination of “Title and Abstract” to conduct patent retrieval [64,77,80,117].

- RQ6: Widely Reported Challenges

The surveyed studies have reported encountering numerous challenges in conducting patent prior art or novelty searches. Among these challenges, “inconsistent document structure and formats” emerged as a prominent issue, highlighted by at least ten studies as a significant obstacle to effective patent retrieval. Similarly, another eight studies identified “sophisticated patent language and vocabulary” as a major hurdle, complicating the search and analysis process. These findings underscore the complexities inherent in patent searches, stemming from both the diversity of document presentations and the specialized nature of patent terminology. These challenges emphasize the need for a specialized NLP model for patent documents that can decode the sophisticated and nuanced language of patent documents.

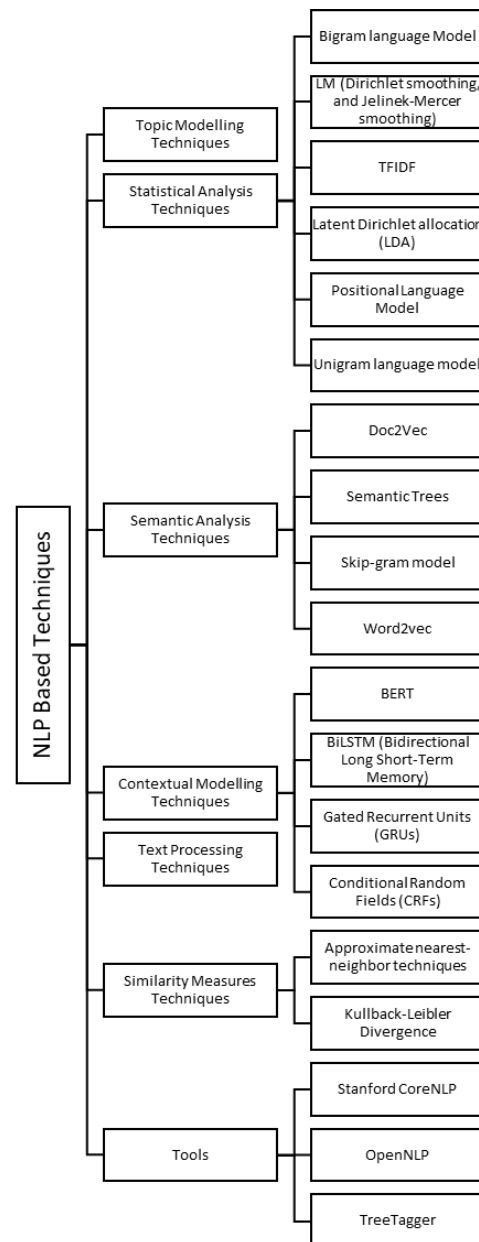


Figure 14. Classification of NLP-based techniques.

4.5. Limitations and Future Directions

This systematic literature review (SLR) is subject to several limitations that should be considered when interpreting its findings. Firstly, the temporal scope of this review is confined to the past decade. While this allows for a focus on recent advancements and trends, it may omit relevant insights and foundational work published before this period, possibly skewing the understanding of the evolution and development of patent retrieval techniques. Secondly, the review is specifically concentrated on patent prior art retrieval. This focus provides depth in one area but limits the breadth of the investigation, potentially overlooking broader aspects of patent retrieval such as infringement checks or patent validity analysis, which could provide a more holistic view of the field. Thirdly, the limitation of sourcing only from journal articles and conference papers might miss significant insights found in other forms of literature, such as book chapters, industry reports, and technical reports. These sources often contain valuable practical applications and case studies that could offer a different perspective or complement the academic viewpoints presented.

Moreover, the review exclusively considers English-language articles, thereby restricting the diversity of viewpoints and potentially missing important contributions from non-English-speaking regions. This language barrier may introduce a bias towards English-speaking researchers' perspectives and methodologies, possibly overlooking innovative approaches developed in other linguistic contexts. Additionally, the exclusion of non-text-based data might lead to overlooking complex innovations that are not easily described in text form. This exclusion could limit the comprehensiveness of the review, especially in fields where visual data are paramount. Lastly, the omission of documents for which full-text access is not available could result in overlooking critical studies. This reliance on accessible full-text documents might bias the review towards more readily available or popular sources, potentially missing out on pivotal but less accessible research.

Each of these limitations not only frames the current findings but also sets the stage for future research directions. Addressing these limitations in subsequent studies could expand the understanding of patent retrieval practices, offering a more comprehensive and inclusive perspective. Future research could aim to include a wider range of sources, extend the temporal coverage, incorporate multilingual studies, and consider non-textual data to enhance the richness and applicability of the findings.

The introduction of Generative AI in patent retrieval offers a promising avenue to address some of these limitations by enhancing search accuracy, automating patent classification, and providing advanced summarization and translation features. With its sophisticated ability to comprehend the nuances of search queries, Generative AI could revolutionize prior art searches, forecast innovation trends, and enable more dynamic and interactive querying processes. However, the integration of this technology must be approached with caution, carefully considering potential challenges such as bias, data privacy, and the costs associated with implementation.

Future research should also broaden the scope of sources used in patent retrieval. This includes incorporating multilingual and non-textual data to better capture the global and diverse range of innovations. Additionally, while non-patent literature may not traditionally fall within the strict bounds of patent retrieval, its integration is crucial for comprehensive prior art searches to establish novelty. Exploring non-patent literature alongside traditional patent databases enriches the context and depth of searches, potentially uncovering prior art that patent documents alone might miss. Furthermore, examining different sections of patent documents, such as claims and descriptions, can provide a richer and more comprehensive dataset for analysis. Expanding the variety of databases explored in research, from well-established patent databases to newer or less utilized sources, can also enhance the depth and breadth of retrieval results. By embracing a wider array of information sources, researchers can develop more robust systems for patent analysis that better reflect the complexities and nuances of innovation across different industries and regions.

Further, embracing interdisciplinary approaches that utilize different technological advances can significantly enrich the field of patent retrieval. Techniques from data analytics, information retrieval, and artificial intelligence could be synergistically combined to develop more robust systems for managing and analyzing patent information. These advancements aim to make patent retrieval more comprehensive, inclusive, and effective, enabling researchers to navigate the complexities of global innovations more efficiently.

This forward-looking approach is designed not only to address the identified limitations but also to harness cutting-edge technologies and methodologies to expand the boundaries of what is currently possible in patent retrieval research. By adapting to and integrating these innovations, the field can evolve to meet the challenges of an increasingly complex intellectual property landscape.

5. Conclusions

Patent prior art retrieval is essential for maintaining the robustness and integrity of the patent system. It ensures that new patents are truly innovative and non-obvious by comparing them against existing inventions. Effective retrieval of prior art helps protect

intellectual property rights and prevents costly legal disputes over patent infringement. Moreover, understanding the landscape of prior art supports strategic R&D investments, guiding inventors and companies away from already patented technologies.

This systematic literature review (SLR) offers a comprehensive examination of state-of-the-art methodologies and their effectiveness, significant data repositories, frequently utilized patent components, the impacts of semantic search and natural language processing, and the key barriers within the domain of patent prior art retrieval. To the best of our knowledge, this is the first survey of its kind in the existing literature. A total of 78 articles, published between 2013 and 2023, focusing on patent retrieval techniques for novelty and prior art searches were extracted from five main research databases through a rigorous search strategy. These articles were thoroughly and meticulously scrutinized, resulting in the identification of a number of research trends and challenges in the domain.

Traditionally, query construction and graph-based techniques have been the preferred methods among researchers for performing patent retrieval. However, recent trends show a significant shift toward employing Natural Language Processing (NLP)-based approaches, which have notably improved the recall rates of patent retrieval tasks. The Cross-Language Evaluation Forum for Intellectual Property (CLEF-IP) and United States Patent and Trademark Office (USPTO) datasets were prominently used in the majority of the studies. To enhance the results of the patent retrieval tasks, most studies made use of the full patent document, despite the resource burden of considering the whole document. Moreover, inconsistent document structure and formats and sophisticated patent language and vocabulary have been the most widely reported challenges in the surveyed studies.

The insights and trends highlighted in this SLR aim to serve as a foundational resource for researchers and practitioners involved in patent novelty and prior art searches, as well as general patent retrieval. Looking ahead, the application of specialized NLP transformer models, such as BERT for Patents, needs to be explored to further enhance the efficacy of novelty and prior art search processes. Moreover, the use of Generative AI also holds immense potential to transform the process of retrieving patents by improving search accuracy, automating patent classification, and providing advanced summarization and translation features.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/asi7050091/s1>, Section S1: Introduction; Section S2: Supplementary Materials to Section 2; Section S3: Supplementary Materials to Section 3 Review Methodology Background.

Author Contributions: Conceptualization, A.A., A.T. and P.E.A.; methodology, A.A., A.T., P.E.A.; validation, A.A., A.T., L.C.D.S. and P.E.A.; formal analysis, A.A., A.T. and P.E.A.; investigation, A.A. and A.T.; resources, A.A. and P.E.A.; data curation, A.A., A.T., L.C.D.S. and P.E.A.; writing—original draft preparation, A.A., A.T.; writing—review and editing, A.A., A.T. and P.E.A.; visualization, A.A. and A.T.; supervision, L.C.D.S. and P.E.A.; funding acquisition, P.E.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universiti Brunei Darussalam grant number UBD/RSCH/1.3/FICBF(b)/2024/023.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fisher, W. Intellectual Property and Innovation: Theoretical, Empirical and Historical Perspectives. In Proceedings of the Programme Seminar on Intellectual Property and Innovation in the Knowledge-Based Economy, The Hague, The Netherlands, 2 May 2001. Available online: <https://cyber.harvard.edu/people/tfisher/Innovation.pdf> (accessed on 23 January 2024).
2. Hallenborg, L.; Ceccagnoli, M.; Clendenin, M. Chapter 3: Intellectual property protection in the global economy. In *Advances in the Study of Entrepreneurship, Innovation, and Economic Growth*; Emerald Group Publishing Limited: Leeds, UK, 2018; Volume 18, pp. 63–116. [CrossRef]
3. Rubilar-Torrealba, R.; Chahuán-Jiménez, K.; de la Fuente-Mella, H. Analysis of the Growth in the Number of Patents Granted and Its Effect over the Level of Growth of the Countries: An Econometric Estimation of the Mixed Model Approach. *Sustainability* **2022**, *14*, 2384. [CrossRef]

4. De Souza Andrade, H.; Urbina LM, S. The Intellectual Property Protection and Commercialization Management Process in a Technology Licensing Office. *Int. J. Adv. Eng. Res. Sci.* **2019**, *6*, 315–331. [CrossRef]
5. Spulber, D.F. How Patents Provide the Foundation of the Market for Inventions. *Northwestern Law Econ. Res.* **2015**, *11*, 271–316. [CrossRef]
6. OECD. *Patents and Innovation: Trends and Policy Challenges*; OECD—Organization for Economic Co-operation and Development: Paris, France, 2004.
7. Othmani, A.; Ben Yedder, N.; Bakari, S. The Cointegration Relationship between Patent, Domestic Investment and Economic Growth in United States of America. MPRA. 2023. Available online: <https://mpra.ub.uni-muenchen.de/id/eprint/118245> (accessed on 18 February 2024).
8. Shalaby, W.; Zadrozny, W. Patent retrieval: A literature review. *Knowl. Inf. Syst.* **2019**, *61*, 631–660. [CrossRef]
9. Risch, J.; Krestel, R. Domain-specific word embeddings for patent classification. *Data Technol. Appl.* **2019**, *53*, 108–122. [CrossRef]
10. Pogiatis, A. NLP: Contextualized Word Embeddings from BERT. 20 March 2019. Available online: <https://towardsdatascience.com/nlp-extract-contextualized-word-embeddings-from-bert-keras-tf-67ef29f60a7b> (accessed on 19 February 2024).
11. Humayun, M.A.; Yassin, H.; Shuja, J.; Alourani, A.; Abas, P.E. A transformer fine-tuning strategy for text dialect identification. *Neural Comput. Appl.* **2023**, *35*, 6115–6124. [CrossRef] [PubMed]
12. Alok Khode, S.J. A Literature Review on Patent Information Retrieval Techniques. *Indian J. Sci. Technol.* **2017**, *10*, 1–13. [CrossRef]
13. Xu, T.; Zhong, Z.; Wang, L.; Ma, J.; Zhang, Z. Methods for the Intellectual Properties Retrieval from Patents. In Proceedings of the 2023 3rd International Conference on Public Management and Intelligent Society (PMIS 2023), Wuhan, China, 24–26 March 2023; pp. 1125–1130. [CrossRef]
14. Takaki, T.; Fujii, A.; Ishikawa, T. Associative document retrieval by query subtopic analysis and its application to invalidity patent search. In Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, Washington, DC, USA, 8–13 November 2004. [CrossRef]
15. Bashir, S.; Rauber, A. Improving Retrievalability of Patents in Prior-Art Search. In Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, 28–31 March 2010; Springer: Berlin/Heidelberg, Germany, 2010; Volume 5993. [CrossRef]
16. Krestel, R.; Chikkamath, R.; Hewel, C.; Risch, J. A Survey on Deep Learning for Patent Analysis. *World Pat. Inf.* **2021**, *65*, 102035. [CrossRef]
17. Rizvi, J. The Importance of a Patent Search. 2024. Available online: <https://thepatentprofessor.com/the-importance-of-a-patent-search/> (accessed on 20 February 2024).
18. Bonino, D.; Ciaramella, A.; Corno, F. Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. *World Pat. Inf.* **2010**, *32*, 30–38. [CrossRef]
19. Kumar, J.L.A. Deep Dive into the Search Function in the Field of Patent (Part 2)—Characteristics of Different Types of Searches. 2020. Available online: <https://www.lexology.com/library/detail.aspx?g=a492d4e0-d08f-4e12-90f1-c355b3402052> (accessed on 21 February 2024).
20. Lupu, M.; Hanbury, A. Patent Retrieval. In *Foundations and Trends® in Information Retrieval*; Now Publishers: Norwell, MA, USA, 2013; Volume 7, pp. 1–97. [CrossRef]
21. Casola, S.; Lavelli, A. Summarization, simplification, and generation: The case of patents. *Expert Syst. Appl.* **2022**, *205*, 117627. [CrossRef]
22. Abbas, A.; Zhang, L.; Khan, S.U. A literature review on the state-of-the-art in patent analysis. *World Pat. Inf.* **2014**, *37*, 3–13. [CrossRef]
23. Bouadjenek, M.R.; Sanner, S.; Ferraro, G. A Study of Query Reformulation for Patent Prior Art Search with Partial Patent Applications. In Proceedings of the ICAIL: International Conference on Artificial Intelligence and Law, San Diego, CA, USA, 8–12 June 2015; pp. 23–32. [CrossRef]
24. Madani, F.; Weber, C. The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis. *World Pat. Inf.* **2016**, *46*, 32–48. [CrossRef]
25. Zhang, L.; Li, L.; Li, T. Patent Mining: A Survey. *ACM SIGKDD Explor. Newsl.* **2015**, *16*, 1–9. [CrossRef]
26. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
27. Jiyun Luo, B.E. Prior Art Search and Its Evaluation. Master's Thesis, Georgetown University, Washington, DC, USA, 2014. Available online: https://repository.library.georgetown.edu/bitstream/handle/10822/709744/Luo_georgetown_0076M_12676.pdf?sequence=1&isAllowed=y (accessed on 25 February 2024).
28. Ali, O.; Abdelbaki, W.; Shrestha, A.; Elbasi, E.; Alryalat MA, A.; Dwivedi, Y.K. A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *J. Innov. Knowl.* **2023**, *8*, 100333. [CrossRef]
29. Keele, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*; EBSE: Durham, UK, 2007.
30. SJR. Scimago Journal & Country Rank. Scopus. Available online: <http://www.scimagojr.com> (accessed on 5 February 2024).
31. Clarivate. Journal Citation Repots. Available online: <https://jcr.clarivate.com/> (accessed on 5 February 2024).
32. Clarivate. Master Journal List. Available online: <https://mjl.clarivate.com/> (accessed on 5 February 2024).

33. Journal Guide. American Journal Experts. Available online: <https://www.journalguide.com/> (accessed on 5 February 2024).
34. CORE. Computing Research and Education, Conference Portal. Available online: <http://portal.core.edu.au/conf-ranks/> (accessed on 5 February 2024).
35. University of Oxford Ranked Conference List. Available online: <http://www.cs.ox.ac.uk/people/michael.wooldridge/conferences.html> (accessed on 5 February 2024).
36. Google Scholar. Available online: <https://scholar.google.com/> (accessed on 5 February 2024).
37. Scopus. Available online: <https://www.scopus.com/> (accessed on 5 February 2024).
38. Singh, D.; Singh, B. Investigating the impact of data normalization on classification performance. *Appl. Soft Comput.* **2020**, *97*, 105524. [CrossRef]
39. Mahdabi, P.; Crestani, F. The effect of citation analysis on query expansion for patent retrieval. *Inf. Retr.* **2013**, *17*, 412–429. Available online: <https://api.semanticscholar.org/CorpusID:254577880> (accessed on 3 March 2024). [CrossRef]
40. Lee, J.; Park, S.; Lee, J. A Fast and Scalable Algorithm for Prior Art Search. *IEEE Access* **2022**, *10*, 7396–7407. [CrossRef]
41. Oh, S.; Lei, Z.; Lee, W.C.; Yen, J. Patent Evaluation Based on Technological Trajectory Revealed in Relevant Prior Patents. In *Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, 13–16 May 2014. Proceedings, Part I 18*; Springer: Cham, Switzerland, 2014.
42. Giachanou, A.; Salampasis, M. IPC Selection Using Collection Selection Algorithms. In *Multidisciplinary Information Retrieval: 7th Information Retrieval Facility Conference, IRFC 2014, Copenhagen, Denmark, 10–12 November 2014, Proceedings 7*; Springer: Cham, Switzerland, 2014.
43. Salampasis, M.; Giachanou, A.; Paltoglou, G. Multilayer Collection Selection and Search of Topically Organized Patents. In *Proceedings of the Integrating IR Technologies for Professional Search Workshop, Moscow, Russia, 24 March 2013*. Available online: <https://api.semanticscholar.org/CorpusID:13514893> (accessed on 3 March 2024).
44. Stamatis, V. End to End Neural Retrieval for Patent Prior Art Search. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, 10–14 April 2022, Proceedings, Part II*; Springer: Cham, Switzerland, 2022.
45. Giachanou, A.; Salampasis, M.; Paltoglou, G. Multilayer source selection as a tool for supporting patent search and classification. *Inf. Retr. J.* **2015**, *18*, 559–585. [CrossRef]
46. Mahdabi, P.; Crestani, F. Query-Driven Mining of Citation Networks for Patent Citation Retrieval and Recommendation. In *Proceedings of the CIKM '14: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, Shanghai, China, 3–7 November 2014*; pp. 1659–1668. [CrossRef]
47. Marrara, S.; Pasi, G. Flexibility in Patent Search. In *Proceedings of the Conference of International Fuzzy Systems Association and European Society for Fuzzy Logic and Technology, Gijon, Spain, 30 June–3 July 2015*. Available online: <https://api.semanticscholar.org/CorpusID:34006451> (accessed on 10 March 2024).
48. Rattinger, A.; Goff, J.M.L.; Guetl, C. Semantic and Topological Graphs for Patent Retrieval. In *Proceedings of the 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 22–25 October 2019*.
49. Albarede, L.; Mulhem, P.; Goeuriot, L.; Le Pape-Gardeux, C.; Marie, S.; Chardin-Segui, T. Passage Retrieval on Structured Documents Using Graph Attention Networks. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, 10–14 April 2022, Proceedings, Part II*; Springer: Cham, Switzerland, 2022.
50. Song, J.J.; Lee, W. Relevance maximization for high-recall retrieval problem: Finding all needles in a haystack. *J. Supercomput.* **2017**, *76*, 7734–7757. [CrossRef]
51. Song, J.J.; Lee, W.; Afshar, J. Retrieving patents with inverse patent category frequency. In *Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp), Hong Kong, China, 18–20 January 2016*.
52. Trappey, A.J.C.; Trappey, C.V.; Wu, J.-L.; Wang, J.W.C. Intelligent compilation of patent summaries using machine learning and natural language processing techniques. *Adv. Eng. Inform.* **2020**, *43*, 101027. [CrossRef]
53. Thang Duong, C.; Percia David, D.; Dolamic, L.; Mermoud, A.; Lenders, V.; Aberer, K. From Scattered Sources to Comprehensive Technology Landscape: A Recommendation-based Retrieval Approach. *World Pat. Inf.* **2023**, *73*, 102198. [CrossRef]
54. Krstovski, K.; Smith, D.A.; Wallach, H.M.; McGregor, A. Efficient Nearest-Neighbor Search in the Probability Simplex. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval, Copenhagen, Denmark, 29 September–2 October 2013*. [CrossRef]
55. Song, J.; Lee, W. High Recall-Low Cost Model for Patent Retrieval. In *Proceedings of the 2015 International Conference on Big Data Applications and Services, Jeju, Republic of Korea, 20–23 October 2015*. [CrossRef]
56. Kamateri, E.; Stamatis, V.; Diamantaras, K.; Salampasis, M. Automated Single-Label Patent Classification using Ensemble Classifiers. In *Proceedings of the 2022 14th International Conference on Machine Learning and Computing, Guangzhou, China, 18–21 February 2022*. [CrossRef]
57. Feng, F.; Li, X. Application of improved chaos theory genetic multi feature matching algorithm in patent retrieval. *J. Ambient Intell. Humaniz. Comput.* **2018**, 1–9. [CrossRef]
58. Bashir, S. Combining pre-retrieval query quality predictors using genetic programming. *Appl. Intell.* **2014**, *40*, 525–535. [CrossRef]
59. Phan, C.-P.; Nguyen, H.-Q.; Nguyen, T.-T. Ontology-based heuristic patent search. *Int. J. Web Inf. Syst.* **2019**, *15*, 258–284. [CrossRef]

60. Girthana, K.; Swamynathan, S. Query Oriented Extractive-Abstractive Summarization System (QEASS). In Proceedings of the CODS-COMAD '19: ACM India Joint International Conference on Data Science and Management of Data, Kolkata, India, 3–5 January 2019.
61. Kumaravel, G.; Sankaranarayanan, S. PQPS: Prior-Art Query-Based Patent Summarizer Using RBM and Bi-LSTM. *Mob. Inf. Syst.* **2021**, *2021*, 2497770. [CrossRef]
62. Wu, H.; Shen, G.; Lin, X.; Li, M.; Zhang, B.; Li, C.Z. Screening patents of ICT in construction using deep learning and NLP techniques. *Eng. Constr. Archit. Manag.* **2020**, *27*, 1891–1912. [CrossRef]
63. Supraja, A.M.; Archana, S.; Suvetha, S.; Geetha, T.V. Patent search and trend analysis. In Proceedings of the 2015 IEEE International Advance Computing Conference (IACC), Bangalore, India, 12–13 June 2015.
64. Sun, D. Multi-Granularity Information Expression Application on Patent Text Clustering. In Proceedings of the 2021 8th International Conference on Dependable Systems and Their Applications (DSA), Yinchuan, China, 5–6 August 2021.
65. Deng, N.; Lin, S.; Xiong, C.; Li, D. A Clustering Algorithm of Four Character Medicine Effect Phrases in TCM Patents. In Proceedings of the 2018 8th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 15–17 June 2018.
66. Zihayat, M.; Etwaroo, R. A non-factoid question answering system for prior art search. *Expert Syst. Appl.* **2021**, *177*, 114910. [CrossRef]
67. Siddharth, L.; Li, G.; Luo, J. Enhancing Patent Retrieval using Text and Knowledge Graph Embeddings: A Technical Note. *J. Eng. Des.* **2022**, *33*, 670–683. [CrossRef]
68. Deng, W.; Huang, X.; Zhu, P. Facilitating Technology Transfer by Patent Knowledge Graph. In Proceedings of the Hawaii International Conference on System Sciences, Maui, HI, USA, 8–11 January 2019. Available online: <https://api.semanticscholar.org/CorpusID:102352345> (accessed on 20 February 2024).
69. Risch, J.; Krestel, R. Learning Patent Speak: Investigating Domain-Specific Word Embeddings. In Proceedings of 2018 Thirteenth International Conference on Digital Information Management (ICDIM), Berlin, Germany, 24–26 September 2018.
70. Rattinger, A.; Le Goff, J.M.; Meersman, R.; Guetl, C. Semantic and Topological Patent Graphs: Analysis of Retrieval and Community Structure. In Proceedings of the International Conference on Social Networks Analysis, Management and Security (SNAMS), Valencia, Spain, 15–18 October 2018.
71. Kravets, A.G.; Korobkin, D.M.; Dykov, M.A. E-patent examiner: Two-steps approach for patents prior-art retrieval. In Proceedings of the 2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA), Corfu, Greece, 6–8 July 2015.
72. Helmers, L.; Horn, F.; Biegler, F.; Oppermann, T.; Müller, K.R. Automating the search for a patent's prior art with a full text similarity search. *PLoS ONE* **2019**, *14*, e0212103. [CrossRef]
73. Wang, F.; Lin, L. Query construction based on concept importance for effective patent retrieval. In Proceedings of the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Zhangjiajie, China, 15–17 August 2015.
74. Mahdabi, P.; Gerani, S.; Huang, J.X.; Crestani, F. Leveraging conceptual lexicon: Query disambiguation using proximity information for patent retrieval. In Proceedings of the SIGIR '13: 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 28 July–1 August 2013.
75. Xu, K.; Lin, H.; Lin, Y.; Xu, B.; Yang, L.; Zhang, S. Patent Retrieval Based on Multiple Information Resources. In *Information Retrieval Technology: 12th Asia Information Retrieval Societies Conference, AIRS 2016, Beijing, China, 30 November–2 December 2016, Proceedings 12*; Springer: Cham, Switzerland, 2016.
76. Sharma, P.; Tripathi, R.; Tripathi, R.C. Finding Similar Patents through Semantic Query Expansion. *Procedia Comput. Sci.* **2015**, *54*, 390–395. [CrossRef]
77. Sharma, P.; Tripathi, R.; Singh, V.K.; Tripathi, R.C. Automated patents search through semantic similarity. In Proceedings of the 2015 International Conference on Computer, Communication and Control (IC4), Indore, India, 10–12 September 2015.
78. Mahdabi, P.; Crestani, F. Patent Query Formulation by Synthesizing Multiple Sources of Relevance Evidence. *ACM Trans. Inf. Syst.* **2014**, *32*, 1–30. [CrossRef]
79. Far, M.G.; Sanner, S.; Bouadjene, M.R.; Ferraro, G.; Hawking, D. On Term Selection Techniques for Patent Prior Art Search. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015. [CrossRef]
80. Sharma, P.; Tripathi, R.; Tripathi, R.C. Finding similar patents through semantic expansion. In Proceedings of the 2016 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 7–9 January 2016.
81. Fu, T.Y.; Lei, Z.; Lee, W.C. Patent Citation Recommendation for Examiners. In Proceedings of the IEEE International Conference on Data Mining, Atlantic City, NJ, USA, 14–17 November 2015.
82. Albarede, L.; Mulhem, P.; Goeuriot, L.; Le Pape-Gardeux, C.; Marie, S.; Chardin-Segui, T. Passage retrieval in context: Experiments on Patents. In Proceedings of the CORIA'21, Virtual, 15 April 2021.
83. Andersson, L.; Lupu, M.; Palotti, J.; Hanbury, A.; Rauber, A. When is the Time Ripe for Natural Language Processing for Patent Passage Retrieval? In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 2016. [CrossRef]
84. Feng, W.; Lanfen, L.; Shuai, Y.; Xiaowei, Z. A semantic query expansion-based patent retrieval approach. In Proceedings of the 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Shenyang, China, 23–25 July 2013.

85. Wang, F.; Lin, L. Exploiting semantic knowledge base for patent retrieval. In Proceedings of the 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, China, 29–31 July 2017.
86. Tablan, V.; Bontcheva, K.; Roberts, I.; Cunningham, H. Mimir: An open-source semantic search framework for interactive information seeking and discovery. *J. Web Semant.* **2015**, *30*, 52–68. [\[CrossRef\]](#)
87. Andersson, L.; Mahdabi, P.; Hanbury, A.; Rauber, A. Exploring Patent Passage Retrieval Using Nouns Phrases. In *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, 24–27 March 2013. Proceedings 35*; Springer: Berlin/Heidelberg, Germany, 2013.
88. Bertram, J.; Mandl, T. Ambiguity in patent vocabulary: Experiments with clarity scores for claims and descriptions. In Proceedings of the 2017 9th International Conference on Knowledge and Smart Technology (KST), Chonburi, Thailand, 1–4 February 2017.
89. Wang, F.; Qian, T.; Liu, B.; Peng, Z. Patent expanded retrieval via word embedding under composite-domain perspectives. *Front. Comput. Sci.* **2019**, *13*, 1048–1061. [\[CrossRef\]](#)
90. Lin, F.-R.; Chen, K.-R.; Lin, S.-Y. A Hybrid Patent Prior Art Retrieval Approach Using Claim Structure and Description. In *Proceedings of the 8th International Conference on Knowledge Management in Organizations*; Springer: Dordrecht, The Netherlands, 2014.
91. Korobkin, D.; Fomenkov, S.; Kravets, A.; Kolesnikov, S. Methods of Statistical and Semantic Patent Analysis. In *Creativity in Intelligent Technologies and Data Science: Second Conference, CIT&DS 2017, Volgograd, Russia, 12–14 September 2017, Proceedings 2*; Springer: Cham, Switzerland, 2017. [\[CrossRef\]](#)
92. Taduri, S.; Law, K.H.; Kesan, J.P.; Sriram, R.D. Utilization of Bio-Ontologies for Enhancing Patent Information Retrieval. In Proceedings of the Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 15–19 July 2019.
93. Law, K.H.; Taduri, S.; Lau, G.T.; Kesan, J.P. An Ontology-Based Approach for Retrieving Information from Disparate Sectors in Government: The Patent System as an Exemplar. In Proceedings of the Conference on System Sciences (HICSS), Kauai, HI, USA, 5–8 January 2015.
94. Zhou, D.; Liu, J.; Zhang, S. Query Generation Techniques for Patent Prior-Art Search in Multiple Languages. In *Natural Language Processing and Chinese Computing: Second CCF Conference, NLPCC 2013, Chongqing, China, 15–19 November 2013, Proceedings 2*; Springer: Berlin/Heidelberg, Germany, 2013.
95. Saraswat, N.; Verma, I.; Gupta, V. Catch-phrase based Document Representation for Improved Prior Art Search. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, Kolkata, India, 3–5 January 2019. [\[CrossRef\]](#)
96. Feng, L.; Peng, Z.; Liu, B.; Che, D. Finding Novel Patents Based on Patent Association. In *Web-Age Information Management: 15th International Conference, WAIM 2014, Macau, China, 16–18 June 2014. Proceedings 15*; Springer: Cham, Switzerland, 2014; Volume 8485. [\[CrossRef\]](#)
97. Hofstätter, S.; Rekabsaz, N.; Lupu, M.; Eickhoff, C.; Hanbury, A. Enriching Word Embeddings for Patent Retrieval with Global Context. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, 14–18 April 2019, Proceedings, Part I 41*; Springer: Cham, Switzerland, 2019; pp. 810–818. [\[CrossRef\]](#)
98. Lagus, J.; Loppi, N.; Klami, A. Second-order Document Similarity Metrics for Transformers. In Proceedings of the International Conference on Natural Language and Speech Processing, Virtual, 16–17 December 2022.
99. Kravets, A.G.; Mironenko, A.G.; Nazarov, S.S.; Kravets, A.D. Patent Application Text Pre-processing for Patent Examination Procedure. In Proceedings of the First Conference on Creativity in Intelligent Technologies and Data Science, CIT&DS 2015, Volgograd, Russia, 15–17 September 2015; pp. 105–114.
100. Powers, D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Mach. Learn. Technol.* **2008**, *2*.
101. Kundu, R. F1 Score in Machine Learning: Intro & Calculation. Machine Learning. 16 December 2022. Available online: <https://www.v7labs.com/blog/f1-score-guide> (accessed on 16 February 2024).
102. Otten, N.V. Mean Average Precision Made Simple [Complete Guide]. 14 September 2023. Available online: <https://spotintelligence.com/2023/09/14/mean-average-precision/> (accessed on 16 February 2024).
103. Gaurav, P. Evaluating Information Retrieval Models: A Comprehensive Guide to Performance Metrics. 2023. Available online: <https://medium.com/@prateekgaurav/evaluating-information-retrieval-models-a-comprehensive-guide-to-performance-metrics-78aadac73b4> (accessed on 10 February 2024).
104. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: New York, NY, USA, 2008.
105. Zhang, Y.; Li, S.; Chen, X.; Qian, F.; Zhao, S.; Zhu, S.; Wang, Y. Semantic Based Heterogeneous Information Network Embedding for Patent Citation Recommendation. In Proceedings of the 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE), Beijing, China, 23–25 October 2020.
106. Althammer, S.; Hofstätter, S.; Hanbury, A. Cross-domain Retrieval in the Legal and Patent Domains: A Reproducibility Study. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, 28 March–1 April 2021, Proceedings, Part II 43*; Springer: Cham, Switzerland, 2021.
107. Khode, A.; Jambhorkar, S. Effect of Technical Domains and Patent Structure on Patent Information Retrieval. *Int. J. Eng. Adv. Technol.* **2019**, *9*, 6067–6074. [\[CrossRef\]](#)
108. Song, J.J.; Lee, W.; Afshar, J. An effective High Recall Retrieval method. *Data Knowl. Eng.* **2019**, *123*, 101603. [\[CrossRef\]](#)

109. Hu, P.; Huang, M.; Zhu, X. Finding nuggets in patent portfolios: Core patent mining and its applications. *Tsinghua Sci. Technol.* **2013**, *18*, 339–352. [\[CrossRef\]](#)
110. Guarino, G.; Samet, A.; Cavallucci, D. PaTRIZ: A framework for mining TRIZ contradictions in patents. *Expert Syst. Appl.* **2022**, *207*, 117942. [\[CrossRef\]](#)
111. Jiang, P.; Atherton, M.; Sorce, S. Extraction and linking of motivation, specification and structure of inventions for early design use. *J. Eng. Des.* **2023**, *34*, 411–436. [\[CrossRef\]](#)
112. Choi, H.; Oh, S.; Choi, S.; Yoon, J. Innovation Topic Analysis of Technology: The Case of Augmented Reality Patents. *IEEE Access* **2018**, *6*, 16119–16137. [\[CrossRef\]](#)
113. Wang, F.; Lin, L. Domain lexicon-based query expansion for patent retrieval. In Proceedings of the 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Changsha, China, 13–15 August 2016.
114. Lee, W.; Leung, C.K.S.; Song, J.J. Reducing Noises for Recall-Oriented Patent Retrieval. In Proceedings of the 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, Sydney, Australia, 3–5 December 2014.
115. Seo, W.; Kim, N.; Choi, S. Big Data Framework for Analyzing Patents to Support Strategic R&D Planning. In Proceedings of the 2016 IEEE 14th International Conference on Dependable, Autonomic and Secure Computing, 14th International Conference on Pervasive Intelligence and Computing, 2nd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), Auckland, New Zealand, 8–12 August 2016.
116. Pasche, E.; Gobeill, J.; Kreim, O.; Oezdemir-Zaech, F.; Vachon, T.; Lovis, C.; Ruch, P. Development and tuning of an original search engine for patent libraries in medicinal chemistry. *BMC Bioinform.* **2014**, *15* (Suppl. S1), S15. [\[CrossRef\]](#)
117. Eisinger, D.; Tsatsaronis, G.; Bundschuh, M.; Wieneke, U.; Schroeder, M. Automated Patent Categorization and Guided Patent Search using IPC as Inspired by MeSH and PubMed. *J. Biomed. Semant.* **2013**, *4* (Suppl. S1), S3. [\[CrossRef\]](#)
118. Al-Shboul, B.; Myaeng, S.-H. Wikipedia-based query phrase expansion in patent class search. *Inf. Retr.* **2014**, *17*, 430–451. [\[CrossRef\]](#)
119. Shalaby, W.; Zadrozny, W. Toward an Interactive Patent Retrieval Framework based on Distributed Representations. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018. [\[CrossRef\]](#)
120. Kim, Y.; Croft, W. Improving Patent Search by Search Result Diversification. In Proceedings of the 2015 International Conference on The Theory of Information Retrieval, Northampton, MA, USA, 27–30 September 2015. [\[CrossRef\]](#)
121. Ma, J.-H.; Wang, N.-N.; Yao, S.; Wei, Z.-M.; Jin, S. Similar Patent Search Method Based on a Functional Information Fusion. In Proceedings of the 2018 7th International Conference on Software and Computer Applications, Kuantan, Malaysia, 8–10 February 2018. [\[CrossRef\]](#)
122. Kim, J.; Choi, J.; Park, S.; Jang, D. Patent Keyword Extraction for Sustainable Technology Management. *Sustainability* **2018**, *10*, 1287. [\[CrossRef\]](#)
123. Roda, G.; Tait, J.; Piroi, F.; Zenz, V. CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain. In Proceedings of the Workshop of the Cross-Language Evaluation Forum for European Languages, Corfu, Greece, 30 September–2 October 2009; Volume 1175. [\[CrossRef\]](#)
124. Piroi, F. CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. In Proceedings of the CLEF 2010, Padua, Italy, 20–23 September 2010.
125. Piroi, F.; Lupu, M.; Hanbury, A.; Zenz, V. CLEF-IP 2011: Retrieval in the intellectual property domain. In Proceedings of the CLEF 2011, Amsterdam, The Netherlands, 19–22 September 2011.
126. Piroi, F.; Lupu, M.; Hanbury, A.; Magdy, W.; Sexton, A.; Filippov, I. CLEF-IP 2012: Retrieval experiments in the intellectual property domain. In Proceedings of the CEUR Workshop, Melbourne, Australia, 10–12 December 2012; Proceedings 1178.
127. Iwayama, M.; Fujii, A.; Kando, N.; Takano, A. Overview of patent retrieval task at NTCIR-3. In Proceedings of the CL-2003 Workshop on Patent Corpus Processing, Sapporo, Japan, 12 July 2003. [\[CrossRef\]](#)
128. Fujii, A.; Iwayama, M.; Kando, N. Overview of Patent Retrieval Task at NTCIR-4. In Proceedings of the NTCIR-4, Tokyo, Japan, 2–4 June 2004. Available online: <https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/PATENT/NTCIR4-OV-PATENT-FujiiA.pdf> (accessed on 1 March 2024).
129. Fujii, A.; Iwayama, M.; Kando, N. Overview of Patent Retrieval Task at NTCIR-5. In Proceedings of the NTCIR-5, Tokyo, Japan, 6–9 December 2005. Available online: <https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/PATENT/NTCIR5-OV-PATENT-FujiiA-pp.pdf> (accessed on 1 March 2024).
130. Fujii, A.; Iwayama, M.; Kando, N. Overview of the Sixth NTCR Workshop. In Proceedings of the NTCIR-6, Tokyo, Japan, 15–18 May 2007. Available online: <http://ntur.lib.ntu.edu.tw/retrieve/170726/26.pdf> (accessed on 1 March 2024).
131. Lupu, M.; Piroi, F.; Huang, X.; Zhu, J.; Tait, J. Overview of the TREC 2009 chemical IR track. In Proceedings of the TREC 2009, Gaithersburg, MD, USA, 17–20 November 2009.
132. Lupu, M.; Tait, J.; Huang, J.; Zhu, J. TREC-CHEM 2010: Notebook Report. In Proceedings of the TREC 2010, Gaithersburg, MD, USA, 16–19 November 2010; NIST Special Publication, 500-294. Available online: <https://trec.nist.gov/pubs/trec19/papers/CHEM.OVERVIEW.pdf> (accessed on 1 March 2024).
133. Lupu, M.; Gurulingappa, H.; Filippov, I.; Zhao, J.; Fluck, J.; Jacobs, M.; Huang, J.; Tait, J. Overview of the TREC 2011 Chemical IR track. In Proceedings of the TREC 2011, Gaithersburg, MD, USA, 15–18 November 2011.

134. Goldstein, B. Intellectual Property and Technology Transfer. In *Principles and Practice of Clinical Research*, 4th ed.; Academic Press: Cambridge, MA, USA, 2018. Available online: <https://www.sciencedirect.com/topics/economics-econometrics-and-finance/european-patent-office> (accessed on 1 March 2024).
135. EPO. EP Full-Text Data. Available online: <https://www.epo.org/en/searching-for-patents/data/bulk-data-sets/data> (accessed on 1 March 2024).
136. Google. Google Patents Public Data. Available online: <https://console.cloud.google.com/getting-started> (accessed on 1 March 2024).
137. Team, E.s.W.I. CNIPA Online Gazette, Retrieving a Chinese Document as PDF Version from CNIPA's Gazette. Available online: https://link.epo.org/web/cnipa_document_retrieval_chinese_202108_en.pdf (accessed on 2 March 2024).
138. Sharma, P.; Tripathi, R.C. Patent Database: A Methodology of Information Retrieval from PDF. *Int. J. Database Manag. Syst. IJDMS* **2013**, *5*, 9. [CrossRef]
139. FIPS. Retrieving Official Publications. Available online: https://link.epo.org/web/fips_downloading_full_russian_documents_en.pdf (accessed on 2 March 2024).
140. USPTO. Research Datasets. Available online: <https://www.uspto.gov/ip-policy/economic-research/research-datasets> (accessed on 1 March 2024).
141. Liu, Y.; Sun, Y. China traditional Chinese Medicine (TCM) Patent Database. *World Pat. Inf.* **2004**, *26*, 91–96. [CrossRef]
142. Binhuraib, T. Kullback–Leibler (KL) Divergence and Cross-Entropy. 2023. Available online: <https://taha-huraibb99.medium.com/kullback-leibler-kl-divergence-and-cross-entropy-f16a735af0b0> (accessed on 3 March 2024).
143. Smucker, M.; Allan, J. *An Investigation of Dirichlet Prior Smoothing's Performance Advantage*; The University of Massachusetts, The Center for Intelligent Information Retrieval: Amherst, MA, USA, 2005.
144. Magalhaes, J. Language Models, LM Jelinek-Mercer Smoothing and LM Dirichlet Smoothing. Available online: <http://ctp.di.fct.unl.pt/~jmag/ir/slides/a05%20Language%20models.pdf> (accessed on 1 March 2024).
145. Lv, Y.; Zhai, C. Positional Language Models for Information Retrieval. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval 2009, Boston, MA, USA, 19–23 July 2009. [CrossRef]
146. Approximate Nearest Neighbors (ANN). Available online: <https://www.activeloop.ai/resources/glossary/approximate-nearest-neighbors-ann/> (accessed on 1 March 2024).
147. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
148. Khattab, O.; Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *arXiv* **2020**, arXiv:2004.12832.
149. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2020**, arXiv:1910.01108.
150. Hugging Face. ParaBERT. Available online: <https://huggingface.co/luciegaba/ParaBERT> (accessed on 15 February 2024).
151. Sen, A. SBERT: How to Use Sentence Embeddings to Solve Real-World Problems. 2023. Available online: <https://anirbansen2709.medium.com/sbert-how-to-use-sentence-embeddings-to-solve-real-world-problems-f950aa300c72> (accessed on 16 February 2024).
152. Tsang, S.-H. Review—TinyBERT: Distilling BERT for Natural Language Understanding. TinyBERT, Outperforms MobileBERT, Much Smaller Than BERT. 2022. Available online: <https://sh-tsang.medium.com/review-tinybert-distilling-bert-for-natural-language-understanding-6c49ad03fa94> (accessed on 16 February 2024).
153. Shao, Y.; Mao, J.; Liu, Y.; Ma, W.; Satoh, K.; Zhang, M.; Ma, S. BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval. In Proceedings of the International Joint Conference on Artificial Intelligence, Yokohama, Japan, 11–17 July 2020.
154. Bidirectional LSTM in NLP. Available online: <https://www.geeksforgeeks.org/bidirectional-lstm-in-nlp/> (accessed on 15 February 2024).
155. Mohdsanadzakirizvi, S. A Comprehensive Guide to Build Your Own Language Model in Python. 2023. Available online: <https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-language-model-nlp-python-code/> (accessed on 10 February 2024).
156. Chawla, R. Overview of Conditional Random Fields. 2017. Available online: <https://medium.com/ml2vec/overview-of-conditional-random-fields-68a2a20fa541> (accessed on 10 February 2024).
157. Doc2Vec in NLP. Available online: <https://www.geeksforgeeks.org/doc2vec-in-nlp/> (accessed on 10 February 2024).
158. Otten, N.V. Practical Guide to Doc2Vec & How to Tutorial in Python. Spot Intelligence. 2023. Available online: https://spotintelligence.com/2023/09/06/doc2vec/#What_is_Doc2Vec (accessed on 10 February 2024).
159. Stewart, E. What is a Gated Recurrent Unit (GRU) and How Does it Work? 6 February 2024. Available online: <https://em360tech.com/tech-article/gated-recurrent-unit-gru> (accessed on 10 February 2024).
160. Bakrey, M. All about Latent Dirichlet Allocation (LDA) in NLP. Medium. 2023. Available online: <https://mohamedbakrey094.medium.com/all-about-latent-dirichlet-allocation-lda-in-nlp-6cfa7825034e#:~:text=for%20using%20LDA-,Introduction,collection%20of%20documents%20or%20texts> (accessed on 10 February 2024).
161. geeksforgeeks. POS (Parts-Of-Speech) Tagging in NLP. Available online: <https://www.geeksforgeeks.org/nlp-part-of-speech-default-tagging/> (accessed on 10 February 2024).
162. Zimmerman, V. Getting to Grips with Parse Trees. 2019. Available online: <https://towardsdatascience.com/getting-to-grips-with-parse-trees-6e19e7cd3c3c> (accessed on 10 February 2024).

163. Doshi, S. Skip-Gram: NLP Context Words Prediction Algorithm. Towards Data Science. 2019. Available online: <https://towardsdatascience.com/skip-gram-nlp-context-words-prediction-algorithm-5bbf34f84e0c> (accessed on 10 February 2024).
164. Pradeep. Understanding TF-IDF in NLP: A Comprehensive Guide. Medium. 2023. Available online: <https://medium.com/@er.iit.pradeep09/understanding-tf-idf-in-nlp-a-comprehensive-guide-26707db0cec5> (accessed on 10 February 2024).
165. OpenNLP. Welcome to Apache OpenNLP. Available online: <https://opennlp.apache.org/> (accessed on 10 February 2024).
166. Schmid, H. TreeTagger—A Language Independent Part-of-Speech Tagger. University of Stuttgart. Available online: <https://www.ims.uni-stuttgart.de/en/research/resources/tools/treetagger/> (accessed on 10 February 2024).
167. Pykes, A. What Is Topic Modeling? An Introduction with Examples. Unlock Insights from Unstructured Data with Topic Modelling. Explore Core Concepts, Techniques like LSA & LDA, Practical Examples, and More. 2023. Available online: <https://www.datacamp.com/tutorial/what-is-topic-modeling> (accessed on 10 February 2024).
168. Kumar, A. N-Gram Language Models, Medium, May 28. 2020. Available online: <https://medium.com/analytics-vidhya/n-gram-language-models-9021b4a3b6b> (accessed on 10 February 2024).
169. Karani, D. Introduction to Word Embedding and Word2Vec. 2018. Available online: <https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa> (accessed on 10 February 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.