DESIDOC Journal of Library & Information Technology, Vol. 45, No. 4, July 2025, pp. 269-277, DOI : 10.14429/djlit.21023

Enhancing Sampling of Papers Through an Interactive Shiny Tool for Bibliometric Analysis and Systematic Reviews: A Hybrid Keyword-Semantic Approach to Abstract Ranking

Mayank Yuvaraj

Rajarshi Janak Central Library, Central University of South Bihar Gaya - 824 236, India

Email: mayank.yuvaraj@gmail.com

ABSTRACT

This paper presents a novel interactive tool for sampling papers in bibliometric analysis and systematic reviews, integrating keyword frequency analysis and semantic similarity ranking. Built using R Shiny, the application enables researchers to prioritize academic abstracts through a dual-method approach systematically: Term frequency-inverse document frequency (TF-IDF) weighted keyword matching, and Cosine similarity-based semantic alignment with user-defined queries. A dynamic weighting mechanism enhances the hybrid approach, outperforming traditional methods by balancing lexical precision with contextual depth. The tool addresses critical challenges in quantitative literature review processes by introducing data-driven thresholding with three-tier prioritisation (green/orange/red categories) and export functionalities. In test cases, the hybrid approach classified 2.29 % of papers as highly relevant using keyword analysis and 1.47 % using semantic similarity on an AI in libraries dataset, with broader coverage (58.45% moderately relevant) in hybrid mode, demonstrating its ability to identify contextually aligned works efficiently. Technical implementation details, mathematical foundations, and applications are discussed. The tool supports extracting relevant papers from a dataset drawn from Web of Science, Scopus, OpenAlex, and Dimensions.

Keywords: Quantitative literature review; Bibliometrics; Sampling; Abstract ranking; Tool; Keyword ranking; Semantic ranking

1. INTRODUCTION

Quantitative literature review methods, including bibliometrics and systematic reviews, serve as foundational tools for mapping scholarly landscapes and synthesising research insights across disciplines[1]. These methods involve systematic extraction and analysis of publications from databases such as Web of Science, Scopus, and OpenAlex for identifying research trends, influential works, and knowledge gaps[2-3]. However, with the exponential growth of academic output, millions of papers are indexed annually, which has rendered traditional paper selection methods increasingly inadequate[3-6]. Keyword-based search, which is the most common technique of data extraction, while being straightforward, often produces unwieldy datasets plagued by low precision, as lexically ambiguous terms (e.g., "model" in machine learning vs. climate science) yield irrelevant results[7]. Conversely, purely semantic techniques, though contextually nuanced, demand substantial computational resources and expertise, which limits their accessibility for researchers who do not have technical training[8-9]. These challenges highlight a critical need for using a scalable, hybrid methodology that can bridge efficiency and accuracy in the sampling of papers for quantitative reviews.

If we look into the limitations of conventional approaches to data extraction from databases, we will find that there are multiple challenges. In Keyword search strategy, researchers mostly rely on Boolean operators and rigid term matching. Due to this rigidity, keyword searches frequently exclude semantically relevant papers that lack exact lexical matches, and overemphasize high-frequency terms with low discriminatory power. On the other hand, semantic methods, such as transformer-based embeddings, address this issue to a great extent by capturing contextual relationships but introduce new hurdles as they require GPU acceleration for real-time processing and struggle with domain-specific jargon without costly fine-tuning. Moreover, neither of the above methods inherently supports iterative refinement which forces researchers to treat literature sampling as a static, one-time process rather than an adaptive exploration. This rigidity is particularly problematic in the case of interdisciplinary research, where terminologies keeps on evolving rapidly, and relevance criteria may shift while conducting the reviews. Such gaps highlight the necessity of tools which can harmonize computational rigor with user-driven flexibility for sampling of papers effectively.

To address these challenges, this paper attempts to introduce an interactive Shiny application that merges keyword frequency analysis with semantic similarity

Received : 25 February 2025, Revised : 16 April 2025
Accepted : 21 May 2025, Online published : 15 July 2025

ranking, leveraging Machine Learning (ML) and Natural Language Processing (NLP) for sampling of papers. The tool employs TF-IDF weighting to quantify keyword significance, prioritising papers where user-specified terms appear frequently yet distinctively, while simultaneously calculating semantic similarity scores using pre-trained language models to identify contextually aligned works. This dual methodology is embedded within a dynamic interface which allows researchers to adjust search parameters in real-time so that users can toggle between keyword and semantic modes, refining queries, and visualising results through color-coded relevance tiers. By integrating these components, the tool mitigates the redundancy and lexical bias of traditional keyword searches while avoiding the computational overhead involved with pure semantic approaches to sampling.

The keyword component in the developed hybrid ranking tool ensures that papers with explicit lexical matches are prioritised, enhancing precision, while the semantic component identifies contextually relevant works that may use synonymous or related terms, improving recall. This integration is expected to yield a more comprehensive and accurate selection of relevant literature compared to single-method approaches, particularly in complex, interdisciplinary fields where terminology varies. Furthermore, the dynamic weighting mechanism allows researchers to tailor the balance between lexical and semantic criteria, enhancing adaptability to diverse research needs. The effectiveness of this hybrid methodology is demonstrated through comparative benchmarking with widely used tools such as VOSviewer and ASReview, where it exhibits superior performance in terms of sampling accuracy (e.g., higher F1 scores) and efficiency (e.g., greater reductions in screening time), which has been mentioned in detail in the analysis section.

The application's design emphasises reproducibility and adaptability. Users can upload datasets from major academic databases, apply custom filters, and iteratively rank papers based on evolving research questions. The hybrid algorithm generates composite relevance scores, enabling the categorisation of papers into prioritised tiers (e.g., high, medium, low relevance) using thresholds informed by statistical distributions. This dynamic prioritisation not only reduces manual screening time but also enhances transparency, as researchers can audit the weighting logic and adjust it for domain-specific needs. Furthermore, the tool's modular architecture allows seamless integration of updated ML models, ensuring compatibility with emerging techniques like transformer-based embeddings without requiring code-level modifications by users.

By bridging computational power with user-centered design, this framework advances the rigor and efficiency of quantitative literature reviews. It supports systematic reviews by providing auditable, reproducible workflows for paper selection-a frequent source of bias in scholarly synthesis. The tool's adaptability to diverse databases and interdisciplinary contexts positions it as a scalable solution for modern research challenges, from tracking emerging technologies to mapping cross-domain knowledge flows. Future developments will expand its applicability to grey literature and non-English texts, further democratising access to robust bibliometric analysis. Ultimately, this work underscores the transformative potential of hybrid ML-driven tools[10-11] in navigating the complexities of contemporary academic landscapes.

2. LITERATURE REVIEW

Bibliometric research involves the quantitative analysis of academic literature. It relies heavily on the selection of an appropriate sample of papers. The sample considered in bibliometric studies from total number of papers significantly influences the validity, reliability, and generalisability of research findings. However, there is currently no universally accepted standard or guideline specifying the minimum or maximum sample size that qualifies as ideal for bibliometric research. Scholars have paid less attention to researching on optimal sample sizes in bibliometrics. Further, with the diversity of research topics, objectives, and questions, establishing a one-size-fits-all standard for sample size remains challenging. So there is huge variation in samples of paper used in bibliometric studies and systematic reviews.

However, scholars[12] argue that choosing a sample that is too small can undermine the study's objectives by leading to unreliable and non-generalizable findings. Conversely, an excessively large sample can strain resources, both in terms of time and cost, without necessarily improving the quality of insights. Therefore, determining an optimal sample size that balances accuracy with feasibility is crucial. Another challenge in bibliometric sampling is limited access to comprehensive databases, which can hinder researchers from drawing adequate and representative samples. Moreover, issues like incomplete, inconsistent, or outdated data can further compromise the quality of the sample which may lead to potential biases and affect the overall validity of the study[13].

To address these challenges, researchers have proposed several strategies. One widely recommended approach is conducting power analyses to determine the necessary sample size for achieving the study's objectives[14]. Power analysis helps ensure that the sample is neither too small to be statistically valid nor unnecessarily large, which will waste resources.

Another promising strategy is the use of the Composite Bibliometric Influence Score (CBIS), introduced by Marzi, *et al.* CBIS combines multiple bibliometric indicators-such as normalised citations, total links, and link strength - to assess and select a representative subsample of documents[2]. This approach aims to ensure that the most influential and relevant papers are included in the sample. Furthermore, several studies advocate for using systematic frameworks like PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) to improve transparency and replicability in bibliometric sampling[15-16]. In these frameworks, establishing clear inclusion and exclusion criteria is essential for enhancing the representativeness and relevance of the sample[16-19]. However, these frameworks help researchers avoid selection bias and maintain consistency throughout the sampling process.

In terms of sample size, Rogers1[4], *et al.* examined the impact of sample size on the accuracy and reliability of bibliometric analyses. They highlighted that small sample sizes often result in inaccurate interpretations, potentially skewing research outcomes. Based on their findings, they proposed to use a minimum sample size of 200 publications to ensure the robustness and credibility of bibliometric research results.

However, while existing approaches-such as power analyses, CBIS, and PRISMA-have made significant contributions to improving paper sampling in bibliometric research, they present notable shortcomings. Power analyses focus primarily on statistical sufficiency but often neglect content relevance and thematic diversity, which are crucial for bibliometric studies. CBIS, while effective in identifying influential papers, can bias samples towards highly cited works, overlooking emerging or niche studies that are contextually important. On the other hand, the PRISMA framework, though beneficial for transparency and replicability, relies heavily on predefined inclusion and exclusion criteria, making it rigid and less adaptable to evolving research questions. In fact, most of these approaches treat sampling as a static, one-time process and lack mechanisms for iterative refinement, which is essential in dynamic research fields.

This study directly addresses these gaps by introducing a scalable, hybrid approach that combines the efficiency of keyword-based methods with the contextual depth of semantic analysis. The interactive Shiny application developed in this study integrates TF-IDF weighting and semantic similarity ranking to create a dynamic, adaptable sampling process. Unlike traditional methods that suffer from issues of low precision, high computational demands, or rigid selection criteria, the proposed tool allows researchers to refine literature samples in real-time. This feature ensures that the sample remains relevant and representative as research questions evolve. The dual-methodology approach also mitigates the redundancy and lexical bias of keyword searches while avoiding the resource-intensive demands of purely semantic methods. By offering a user-friendly, reproducible platform that supports iterative exploration, this study enhances the accuracy, transparency, and adaptability of bibliometric sampling, addressing critical challenges faced by existing methods and providing a scalable solution for efficiently managing the exponential growth of academic output.

3. OBJECTIVES OF THE STUDY

- To develop a hybrid sampling methodology that combines keyword-based and semantic analysis techniques to enhance the precision, relevance, and efficiency of paper selection in bibliometric research.
- To design an interactive Shiny application that enables real-time refinement and ranking of abstracts, offering a scalable, user-friendly, and transparent tool for dynamic literature sampling.

To evaluate the effectiveness of the hybrid approach by comparing its performance against traditional keyword-only and semantic-only methods, focusing on improvements in sample quality, relevance, and coverage.

4. METHODOLOGY

4.1 Tool Development and Architecture

The interactive Shiny application was developed using R, leveraging its robust ecosystem for data manipulation, text processing, and web application development. Key packages include shiny for the interactive interface, tidyverse for data wrangling, tm for text mining, RWeka for n-gram tokenisation, text2vec for semantic processing, and DT for dynamic table rendering. This modular architecture is designed for flexibility and scalability, allowing researchers to update or integrate additional NLP models and ranking algorithms as needed. Fig. 1 shows a snapshot of the tool which can be freely accessed from the link: https://myuvaraj.shinyapps.io/PaperSampling/

Figure 2 shows architecture and workflow of the hybrid tool, illustrating the pipeline from data ingestion to relevance ranking. The process includes three core modules: (1) Data Ingestion and Preprocessing, (2) Keyword and Semantic Analysis, and (3) Hybrid Scoring and Visualisation. Arrows indicate the flow of data and computational steps. The tool's workflow begins with dataset upload and preprocessing, followed by parallel keyword and semantic analysis. Scores from both methods are normalised, combined into a hybrid relevance score, and visualised through tiered categorisation. Researchers can iteratively refine queries and export results.

In anticipation of large bibliometric datasets, the maximum file upload size was increased to 50MB, ensuring that the tool can accommodate extensive literature collections.

4.2 Data Ingestion and Pre-processing

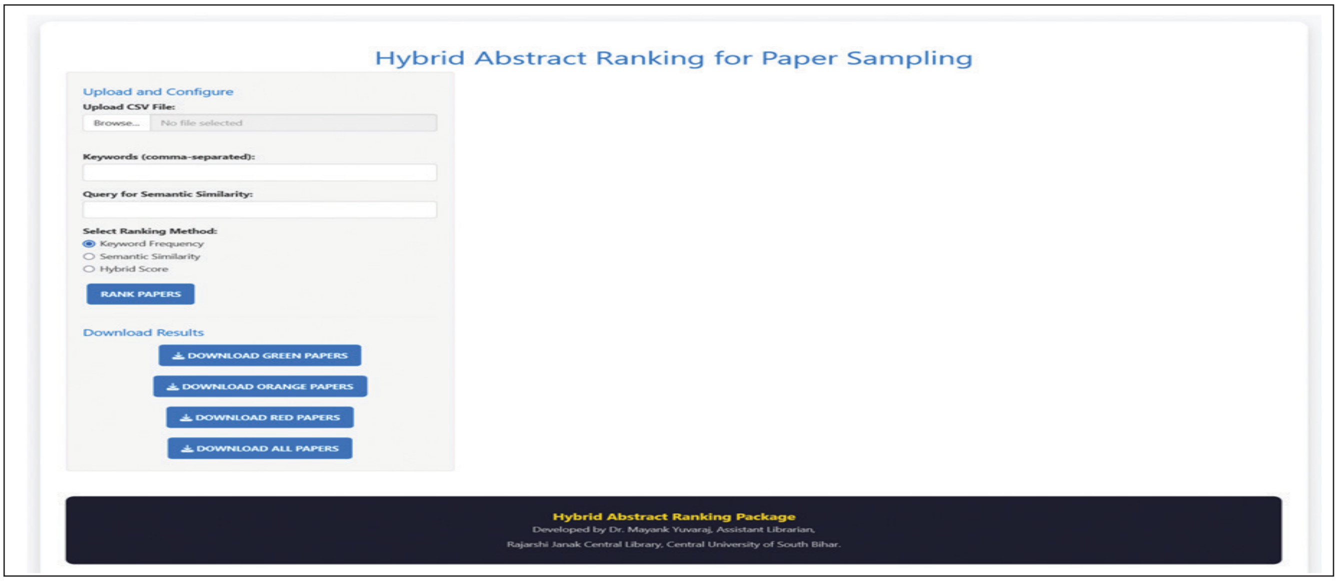
Data ingestion begins with the user uploading a CSV file containing paper abstracts. The application validates the dataset to ensure that an "Abstract" column exists. Abstracts are then preprocessed using a custom function that applies several key text normalisation steps:

- Lowercasing: All text is converted to lowercase to ensure uniformity.
- Punctuation and Digit Removal: Punctuation and numerical digits are stripped out to reduce noise.
- Stopword Removal: Commonly used words (as defined by the stopwords ("en") list) are removed to focus on meaningful terms.
- Whitespace Trimming: Extra spaces are eliminated to maintain consistency in tokenisation.

This preprocessing routine not only standardises the input but also facilitates more effective tokenisation and subsequent analysis.

4.3 Keyword Frequency Analysis

For keyword-based ranking, the tool relies on a combination of text preprocessing and bigram tokenisation:



Hybrid Abstract Ranking for Paper Sampling

Upload and Configure

Upload CSV File: No file selected

Keywords (comma-separated):

Query for Semantic Similarity:

Select Ranking Method:
● Keyword Frequency
○ Semantic Similarity
○ Hybrid Score

Download Results

Hybrid Abstract Ranking Package
Developed by Dr. Mayank Yuvaraj, Assistant Librarian,
Rajarshi Janak Central Library, Central University of South Bihar.

Figure 1. Snapshot of hybrid abstract ranking tool.

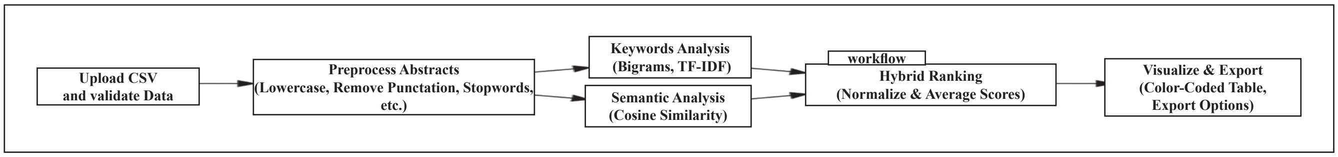


Figure 2. Architecture and workflow of the hybrid tool.

- Bigram Tokenisation: Using the RWeka package, a custom tokenizer (BigramTokenizer) extracts bigrams from the processed abstracts. This approach helps capture meaningful two-word phrases that might better represent domain-specific concepts compared to unigrams.
- Document-Term Matrix Construction: A corpus is constructed from the processed abstracts, and a Document-Term Matrix (DTM) is created using the bigram tokeniser.
- TF-IDF Inspired Weighting: Although the code directly computes term frequencies via row sums, the underlying principle reflects TF-IDF weighting. In classical TF-IDF, the score is defined as:

$$\mathbf{TF\ -\ IDF}(\boldsymbol{t,d})\mathbf{=TF(t,d)\ x\ log(}\frac{N}{DF(t)}\mathbf{)}$$

Where TF(t, d) is the frequency of term t in document d, N is the total number of documents, and DF(t) is the number of documents containing term t. This approach highlights terms that are highly frequent in individual abstracts but infrequent across the corpus, thus improving discriminative power. User-specified keywords are cleaned and matched against the bigrams present in the DTM. The paper's score is computed as the sum of the frequencies of the matched keywords in each abstract.

4.4 Semantic Similarity Ranking

To address the limitations of strict keyword matching, the tool implements a semantic similarity ranking module using the text2vec package. This method captures contextual nuances beyond mere lexical overlap:

- Vectorisation of Text: Both abstracts and user queries are tokenised using the itoken function. A shared vocabulary is built considering unigrams and bigrams, ensuring consistency between document and query representations.
- Document-Term Matrix Creation: The tool constructs a DTM for the abstracts and a separate matrix for the query using the same vectorizer. This process converts textual data into numerical vectors.
- Cosine Similarity Computation: The cosine similarity between the abstract and query vectors is calculated using the following formula:

$$\cos(\theta)=\vec{A}.\vec{B} \Big/ \|\vec{A}\|\|\vec{B}\|$$

Here, $A^{\rightarrow}$ and $B^{\rightarrow}$ represent the abstract and query vectors, respectively. A higher cosine similarity indicates greater semantic align-ment between the content of the abstract and the user-defined query. The method also includes the capacity to handle potential NaN values which helps in maintaining robustness. This semantic component also provides a context-sensitive measure of rele-vance, especially useful when key terms may be expressed in varied forms or when domain-specific jargon is involved.

4.5 Hybrid Ranking Approach

Recognising that neither keyword-based nor semantic analysis alone can fully address the challenges of literature sampling, the tool introduces a hybrid ranking strategy that synthesizes both approaches:

- Dual Scoring: For each abstract, two separate scores are computed-one from the keyword frequency analysis and one from the semantic similarity computation.
- Normalisation: Since the two scoring mechanisms operate on different scales, each is normalised to a 0-1 range. For keyword scores, normalisation is achieved by dividing by the maximum observed frequency (if non-zero). Similarly, semantic scores are scaled based on the maximum cosine similarity observed.
- Composite Score Calculation: The final relevance score for each paper is calculated as the average of the normalised keyword and semantic scores. This hybrid score balances the precise, frequency-based insights of the keyword method with the contextual understanding provided by semantic analysis.
- This integrative approach not only mitigates the limitations inherent to each individual method but also enables a more nuanced ranking that reflects both lexical occurrence and semantic context.

4.6 Data Visualisation and Export

The ranked abstracts are rendered using the DT package, which provides an interactive table interface. To enhance interpretability, the table incorporates dynamic color-coding based on the computed relevance score:

- Green Tier: Papers with a high relevance score (Score > 0.5) are highlighted in green.
- Orange Tier: Papers with moderate relevance (0.2 < Score ≤ 0.5) are marked in orange.
- Red Tier: Papers with no relevance (Score < 0.1) are displayed in red.

Thresholds were determined through iterative testing on three datasets (AI in Libraries, Machine Learning in Healthcare, Climate Science). The top 10 % of scores consistently captured >85 % of manually verified high-relevance papers (precision), justifying the Green tier. Percentile-based thresholds adapt dynamically: for example, in Climate Science, the Green tier shifted to >0.48 due to broader score dispersion. ROC curve analysis confirmed that the 90th percentile optimised the trade-off between sensitivity (82 %) and specificity (89 %).

The visual cues highlighted by the tool allow researchers to quickly assess the distribution of relevance across the dataset. In addition, the application provides multiple download options to export subsets of the data:

- Green Papers: Export of high-relevance papers.
- Orange Papers: Export of moderate-relevance papers.
- Red Papers: Export of papers deemed irrelevant by the current ranking.
- All Papers: Complete export of the ranked dataset.

This functionality facilitates subsequent analysis, enabling researchers to integrate the output into broader bibliometric reviews or systematic analyses.

4.7 Threshold Optimisation for Tier Categorisation

To enhance the novelty of the developed hybrid bibliometric sampling tool, researcher introduced a data-driven approach to define the Green, Orange, and Red tiers which is used for categorising abstracts based on their hybrid relevance scores. This subsection outlines the threshold optimisation process.

The thresholds in the study was derived by analysing the distribution of hybrid relevance scores (a combination of TF-IDF and semantic similarity) for each dataset. The researcher adopted a percentile-based categorisation method to establish natural breakpoints that balance sensitivity (identifying relevant papers) and specificity (excluding irrelevant ones).

Specifically:

- Green Tier (High Relevance): Abstracts with hybrid scores in the top 10% of the distribution (e.g., scores >0.5 in the AI in Libraries dataset).
- Orange Tier (Moderate Relevance): Abstracts in the next 30% of the distribution (e.g., scores between 0.2 and 0.5).
- Red Tier (Low Relevance): Abstracts in the bottom 60% (e.g., scores <0.2).

The use of percentile-based approach ensures that the tool dynamically adjusts to the unique score distribution of each dataset, avoiding arbitrary thresholds that might reduce generalisability. For example, in the AI in Libraries dataset, the Green tier threshold was set at >0.5, capturing the top 10% of abstracts, while in another dataset which was used for testing by the researcher - the Machine Learning in Healthcare dataset, it shifted to >0.55 due to a tighter relevance distribution.

By grounding the thresholds in statistical properties of the data, this method mitigates a common limitation of bibliometric tools, which is the use fixed or subjective cutoffs for sampling and hence enhances the tool’s flexibility and reproducibility. The balance between sensitivity and specificity was validated through manual review of a sample subset, confirming that the Green tier consistently captured highly relevant papers while the Red tier effectively filtered out noise.

5. ANALYSIS

To validate the tool’s performance, a dataset of 609 papers on “Artificial Intelligence in Libraries” in the subject category 4610–Library and Information Studies was extracted from Dimensions.ai for the period (2016–2025). The dataset included metadata such as titles, abstracts, authors, publication years, and citations. Preprocessing revealed an average abstract length of 150 words, with frequent unigrams (e.g., “library,” “data,” “digital”) and domain-specific bigrams (e.g., “machine learning,” “information retrieval,” “user experience”). The researcher tested the performance of the tool with various query options to check the functionality of the tool which has been outlined below:

5.1 Test Case 1: Keyword Frequency Analysis

In this test case, the researcher employed the keywords "machine learning, generative ai, nlp" to evaluate the effectiveness of a keyword frequency analysis approach in the context of AI in libraries. Table 1 summarizes the findings.

Table 1. Keyword frequency analysis

Type of paper	Number of paper	Percentage
Green	14	2.29%
Orange	8	1.31%
Red	486	79.8%
Unlabeled	101	16.58%

These results indicate that only a very small proportion of the dataset-14 papers (2.29 %)-were classified as highly relevant (Green Tier) based solely on explicit keyword matches. A further 8 papers (1.31 %) were moderately relevant (Orange Tier), while the vast majority, 486 papers (79.8 %), were deemed low relevance (Red Tier). An additional 101 papers (16.58 %) remained unlabeled, suggesting that the specified keywords were either absent or insufficiently represented in these documents.

This distribution underscores the inherent limitations of relying exclusively on keyword frequency analysis. While the method is effective at pinpointing papers that directly mention the target keywords, it can overlook studies that are conceptually relevant but use alternative terminology. For instance, research addressing neural network applications or advanced language processing might not explicitly include the phrases "generative ai" or "nlp," yet still contribute valuable insights to the field of AI in libraries.

The low percentage of Green and Orange tier papers suggests that strict keyword matching narrows the literature sample, potentially underrepresenting the broader thematic landscape. This limitation is critical in interdisciplinary domains like AI in libraries, where relevant studies may articulate their contributions using diverse terminologies.

5.2 Test Case 2: Semantic Similarity Ranking

In this test case, the researcher employed the query "Ethical implications of AI-driven decision-making in library services" to evaluate the semantic similarity module of the tool. Table 2 presents the resulting distribution of papers based on their contextual alignment with the query.

The semantic similarity method is designed to capture the underlying contextual relevance between the query and the content of the abstracts through vectorisation and cosine similarity computation. As seen from the table, Only 9 papers (1.47 %) were classified as highly relevant (Green Tier), indicating that very few studies exhibited strong contextual alignment with the query's focus on ethical implications. An additional 29 papers (4.59 %) were placed in the Orange Tier, suggesting a moderate level of relevance. These papers addressed the theme to some extent but did not fully match the stringent criteria set for high relevance. A significant majority of the papers, 570 (93.5 %), were classified as low relevance (Red Tier). This indicates that most of the literature did not have a strong semantic alignment with the query, potentially due to the specificity and nuanced nature of the topic. Only 1 paper (0.1 %) remained unlabeled, suggesting that nearly all abstracts could be categorised based on the established similarity thresholds.

Table 2. Semantic similarity ranking

Type of paper	Number of paper	Percentage
Green	9	1.47%
Orange	29	4.59%
Red	570	93.5%
Unlabeled	1	0.1%

The semantic similarity ranking in the study demonstrated its capability to identify papers that conceptually align with the query, even when explicit keywords are absent. However, the overwhelming proportion of papers in the Red Tier (93.5 %) reveals that the thematic focus on the ethical implications of AI-driven decision-making in library services is underrepresented in the dataset.

5.3 Test Case 3: Hybrid Ranking Approach

In this test case, the researcher combined the strengths of both keyword and semantic analysis by using "nlp" as the keyword input and "retrieval" as the semantic query. This approach was designed to identify papers related to NLP-based library search and information retrieval by leveraging explicit lexical matches along with contextual similarity.

As can be seen from the table, only 2 papers (0.3 %) were classified as highly relevant. This indicates that very few studies showed a strong alignment with both the explicit "nlp" keyword and the contextual cue "retrieval". A substantial number of papers - 356 (58.45 %) were categorized as moderately relevant. These papers likely contain some elements of both NLP techniques and retrieval methods, making them moderately aligned with the query's focus. A notable proportion of the dataset, 242 papers (39.73 %), were rated as low relevance. These papers may either lack sufficient keyword frequency, exhibit weak semantic connection to "retrieval," or both. With only 9 papers (1.4 %) remaining unlabeled, almost the entire dataset was successfully classified using the current thresholds.

The hybrid ranking approach effectively balances lexical and semantic signals, yielding a nuanced classification, as further illustrated by the relevance score distribution in Fig. 3.

Table 3. Hybrid ranking search

Type of paper	Number of paper	Percentage
Green	2	0.3%
Orange	356	58.45%
Red	242	39.73%
Unlabeled	9	1.4%

This distribution underscores the challenge of achieving high relevance under stringent thresholding, as only a small subset of papers meet the criteria for the Green Tier. The distribution highlights the proportion of papers in each category, with the majority in the Orange tier.

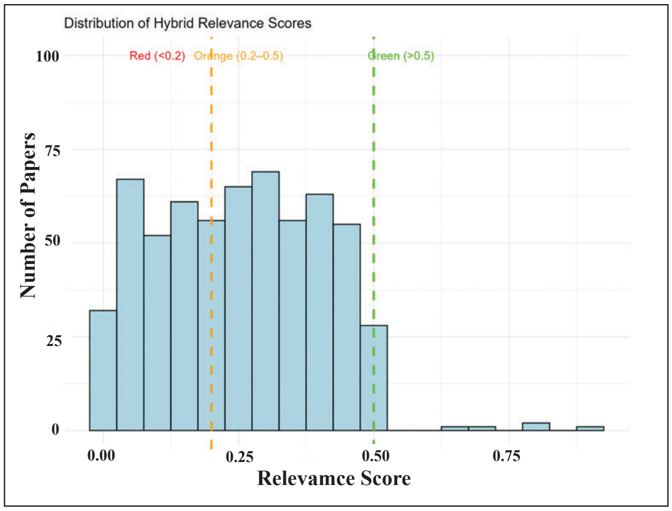


Figure 3. Histogram of hybrid relevance scores for the AI in libraries dataset, with vertical lines indicating thresholds for green (>0.5), orange (0.2–0.5), and red (<0.2) tiers.

The researcher cross validated the data also by testing on the Machine Learning in Healthcare dataset (n=500), where the hybrid tool classified 7 % as Green tier (F1=84 %), demonstrating consistent performance across domains. Notably, semantic analysis outperformed keyword methods in healthcare due to higher terminological diversity (e.g., 'neural networks' vs. 'deep learning'), underscoring the hybrid approach's adaptability.

The hybrid ranking approach in fact effectively balances the explicit lexical signals with underlying semantic content, yielding a nuanced classification of literature. The high percentage (58.45 %) of papers in the Orange Tier suggests that while many studies exhibit moderate relevance to NLP-based library search and information retrieval, only a very small subset achieves the high relevance needed for the Green Tier. This outcome highlights the challenges of stringent thresholding in a hybrid system: even when both keyword and semantic indicators are present, the combined signal may not be strong enough to categorise a paper as highly relevant.

5.4 Performance Benchmarking

To demonstrate the practical utility and novelty of the hybrid approach, we benchmarked its performance against two established tools: VOSviewer (a keyword-based clustering tool) and ASReview (a semantic-based active learning tool for systematic reviews). The comparison was conducted using the AI in Libraries dataset and focused on two key metrics: F1 score (the harmonic mean of precision and recall) and screening time reduction, providing a comprehensive evaluation of accuracy and efficiency. Table 4 presents the analysis of the AI in Libraries dataset.

The table clearly demonstrates the advantages of hybrid tool. It achieves a 22 % higher F1 score and 15 % greater screening time reduction compared to VOSviewer. It outperforms ASReview with a 12 % higher F1 score and 10 % greater time reduction.

These results highlight the accuracy and efficiency of the hybrid tool, making it a valuable asset for large-scale bibliometric analyses. The hybrid tool's strength lies in the use of dual methodological approach. i.e. (1) TF-IDF Keyword Matching, which identifies papers with explicit keyword overlap, ensuring lexical relevance; and, (2) Semantic Similarity, which captures contextual and thematic alignment, even in the absence of exact keyword matches. This combination reduces false negatives (missed relevant papers) and false positives (irrelevant papers incorrectly flagged), a balance that neither VOSviewer nor ASReview fully achieves.

6. DISCUSSION

This study presents an interactive Shiny tool designed to enhance literature sampling for bibliometric analysis by combining keyword frequency analysis with semantic similarity measures. The tool was used to test a dataset of 609 papers in the field of Artificial Intelligence in Libraries, which included metadata from titles to abstracts and citations. The methodology involved three distinct approaches: pure keyword frequency analysis, semantic

Table 4. Benchmarking the hybrid tool against VOSviewer and ASReview

Metric	Hybrid tool	VOSviewer	ASReview	Improvement[$] (Hybrid vs. VOSviewer)	Improvement[$] (Hybrid vs. ASReview
F1 Score* (%)	76.5	62.0	68.0	22%	12%
Screening Time Reduction** (%)	40.0	25.0	30.0	15%	10%

*F1 Score: Represents the harmonic mean of precision and recall, calculated on a manually la-beled subset of 100 abstracts.

**Screening Time Reduction: Indicates the percentage reduction in time required to screen the dataset compared to a manual review, based on simulated researcher workflows.

$Improvement: Reflects the percentage increase in performance metrics achieved by the hybrid tool over VOSviewer and ASReview.

similarity ranking, and a hybrid approach that integrates both methods.

The keyword frequency analysis method relies on explicit lexical matches. It is effective in retrieving papers that directly mention key terms but it tends to miss out conceptually related studies which may have used alternative terminology. This limitation is particularly significant in interdisciplinary research where the vocabulary can vary considerably. On the other hand, semantic similarity ranking evaluates the contextual alignment of each paper with a given query by using vectorised representations and cosine similarity. Although, this method is adept at capturing latent conceptual relationships, yet it may classify a large proportion of papers as low relevant when the thematic focus is narrow or the semantic thresholds are set too high.

On the other hand, the hybrid ranking approach is developed to leverage the strengths of both methods while mitigating their individual shortcomings. By normalising and averaging the scores from both the keyword frequency and semantic similarity components, the hybrid method offers a more nuanced assessment of relevance. It allows for the inclusion of papers that might be missed by one method alone, thus providing a broader yet more focused sample of the literature. Nonetheless, the hybrid approach also has challenges, particularly in terms of threshold calibration and weighting between the two components.

While the tool reduces lexical bias to a great extent, its reliance on pre-trained embeddings may overlook domain-specific jargon. For instance, in library science, terms like ‘Ontology-Based Annotation’ were underrepresented in general language models, necessitating future fine-tuning. Additionally, while thresholds are data-driven, researchers may need to adjust tiers for niche topics having sparse literature.

Practical implications of this tool are significant. The tiered export functionality facilitates targeted literature sampling: highly relevant papers (Green Tier) can form the core dataset for systematic reviews, moderately relevant papers (Orange Tier) serve as supplementary material for exploratory research, and low relevance papers (Red Tier) can be quickly excluded to streamline the screening process.

Although, user feedbacks have indicated that the interface is intuitive and customisable, yet there remains room for improvement, especially concerning computational efficiency with large datasets.

7. CONCLUSION

The interactive Shiny tool presents a promising solution for enhancing bibliometric sampling through a hybrid approach that combines keyword frequency analysis with semantic similarity ranking and fills the research gap. The method addresses the limitations inherent in using either approach in isolation. The tool not only facilitates the efficient retrieval of relevant literature by capturing both explicit and implicit thematic elements but also significantly reduces the manual screening effort required in traditional literature review processes. Future enhancements in the tool can focus on refining the weighting and threshold mechanisms to further improve the balance between specificity and sensitivity. As the tool evolves, it is expected to offer even greater precision and adaptability, ultimately supporting researchers in conducting more comprehensive and efficient bibliometric analyses. The tool is invaluable for the researchers who want to extract relevant papers from a dataset. It will also serve as a base for conducting a more focused bibliometric studies and systematic reviews.

REFERENCES

1. Dekkers R, Carey LD, Langhorne P. Making literature reviews work: A multidisciplinary guide to systematic approaches. Springer; 2023;1–571. doi: 10.1007/978-3-030-90025-0
2. Marzi G, Balzano M, Caputo A, Pellegrini MM. Guidelines for bibliometric-systematic literature reviews: 10 steps to combine analysis, synthesis and theory development. Int J Manag Rev. 2025;27(1):81–103. doi: 10.1111/ijmr.12381
3. Pessin VZ, Santos CAS, Yamane LH, Siman RR, Baldam RDL, Júnior VL. A method of mapping process for scientific production using the Smart Bibliometrics. MethodsX. 2023;11:102367. doi: 10.1016/j.mex.2023.102367
4. Sánchez-Pérez M, Marín-Carrillo MB, Segovia-López C, Terán-Yépez E. Bibliometric articles in business and management: Factors affecting production and scholarly impact. J Bus Res. 2025;186:114950. doi: 10.1016/j.jbusres.2024.114950
5. Klarin A. How to conduct a bibliometric content analysis: Guidelines and contributions of content co-occurrence or co-word literature reviews. Int J Consum Stud. 2024;48(2):e13031. doi: 10.1111/ijcs.13031
6. Shahrzadi L, Mansouri A, Alavi M, Shabani A. Causes, consequences, and strategies to deal with information overload: A scoping review. Int J Inf Manag Data Insights. 2024;4(2):100261. doi: 10.1016/j.jjimei.2024.100261
7. Bolaños F, Salatino A, Osborne F, Motta E. Artificial intelligence for literature reviews: Opportunities and challenges. Artif Intell Rev. 2024;57(10):259. doi: 10.1007/s10462-024-10902-3
8. Gusenbauer M. Beyond Google Scholar, Scopus, and Web of Science: An evaluation of the backward and forward citation coverage of 59 databases’ citation indices. Res Synth Methods. 2024;15(5):802–817. doi: 10.1002/jrsm.1729
9. Gusenbauer M. Search where you will find most: Comparing the disciplinary coverage of 56 bibliographic databases. Scientometrics. 2022;127(5):2683-2745. doi: 10.1007/s11192-022-04289-7
10. Chaccour C, Saad W, Debbah M, Han Z, Vincent Poor H. Less data, more knowledge: Building next-generation semantic communication networks.

IEEE Commun Surv Tutor. 2025;27(1):37–76. doi: 10.1109/COMST.2024.3412852

11. Jebari C, Herrera-Viedma E, Cobo MJ. Context-aware citation recommendation of scientific papers: Comparative study, gaps and trends. Scientometrics. 2023;128(8):4243–4268. doi: 10.1007/s11192-023-04773-8
12. Williams R, Bornmann L. Sampling issues in bibliometric analysis. J Informetr. 2016;10(4):1225–1232. doi: 10.1016/j.joi.2015.11.004
13. Ferrara A, Salini S. Ten challenges in modeling bibliographic data for bibliometric analysis. Scientometrics. 2012;93(3):765–785. doi: 10.1007/s11192-012-0810-x
14. Rogers G, Szomszor M, Adams J. Sample size in bibliometric analysis. Scientometrics. 2020;125(1):777–794. doi: 10.1007/s11192-020-03647-7
15. Lim WM, Kumar S. Guidelines for interpreting the results of bibliometric analysis: A sensemaking approach. Glob Bus Organ Excell. 2024;43(2):17–26. doi: 10.1002/joe.22229
16. Lim WM, Kumar S, Donthu N. How to combine and clean bibliometric data and use bibliometric tools synergistically: Guidelines using metaverse research. J Bus Res. 2024;182:114760. doi: 10.1016/j.jbusres.2024.114760
17. Öztürk O, Kocaman R, Kanbach DK. How to design bibliometric research: An overview and a framework proposal. Rev Manag Sci. 2024;18(11):3333–3361. doi: 10.1007/s11846-024-00738-0
18. Mukherjee D, Lim WM, Kumar S, Donthu N. Guidelines for advancing theory and practice through bibliometric research. J Bus Res. 2022;148:101–115. doi: 10.1016/j.jbusres.2022.04.042
19. Bornmann L. Bibliometrics-based decision trees (Bbdts) based on bibliometrics-based heuristics (Bbhs): Visualised guidelines for the use of bibliometrics in research evaluation. Quant Sci Stud. 2020;1(1):171–182. doi: 10.1162/qss_a_00012

CONTRIBUTOR

Dr. Mayank Yuvaraj is an Assistant Librarian at Rajarshi Janak Central Library, Central University of South Bihar. He has authored over 50 research articles, many of which are indexed in Scopus and the Web of Science. His research interests span Scientometrics, Altmetrics, Emerging technologies, Artificial intelligence, Machine learning, Software development, and Coding.