Article

Automation Applied to the Collection and Generation of Scientific Literature

Nadia Paola Valadez-de la Paz, Jose Antonio Vazquez-Lopez, Aidee Hernandez-Lopez, Jaime Francisco Aviles-Viñas, Jose Luis Navarro-Gonzalez, Alfredo Valentin Reyes-Acosta and Ismael Lopez-Juarez

Automation Applied to the Collection and Generation of Scientific Literature

Nadia Paola Valadez-de la Paz [1], Jose Antonio Vazquez-Lopez [1], Aidee Hernandez-Lopez [2], Jaime Francisco Aviles-Viñas [3], Jose Luis Navarro-Gonzalez [4], Alfredo Valentin Reyes-Acosta [5] and Ismael Lopez-Juarez [3,6,*,†]

[1] Departamento de Ingenieria Industrial, Tecnologico Nacional de Mexico, Instituto Tecnologico de Celaya, Celaya 38010, Mexico; d2303028@itcelaya.edu.mx (N.P.V.-d.l.P.); antonio.vazquez@itcelaya.edu.mx (J.A.V.-L.)
[2] Departamento de Ingenieria Industrial, Universidad del SABES, Campus Celaya, Leon 37234, Mexico; aidee.hernandezl@sabes.edu.mx
[3] Mechatronics Group, Engineering Faculty, Autonomous University of Yucatan, Merida 97302, Mexico; javiles@correo.uady.mx
[4] Departamento de Ingenieria Mecanica, Tecnologico Nacional de Mexico, Instituto Tecnologico de Saltillo, Saltillo 25280, Mexico; jose.ng@saltillo.tecnm.mx
[5] Facultad de Sistemas, Unidad Saltillo, Universidad Autonoma de Coahuila, Saltillo 25350, Mexico; alfredoreyes@uadec.edu.mx
[6] Robotics and Advanced Manufacturing Department, Centre for Research and Advanced Studies (CINVESTAV), Ramos Arizpe 25900, Mexico
* Correspondence: ismael.lopez@cinvestav.edu.mx
† Current address: Engineering Faculty, Autonomous University of Yucatan, Ind No Contaminantes S/N, Col 27, Merida 97302, Mexico.

Abstract: Preliminary activities of searching and selecting relevant articles are crucial in scientific research to determine the state of the art (SOTA) and enhance overall outcomes. While there are automatic tools for keyword extraction, these algorithms are often computationally expensive, storage-intensive, and reliant on institutional subscriptions for metadata retrieval. Most importantly, they still require manual selection of literature. This paper introduces a framework that automates keyword searching in article abstracts to help select relevant literature for the SOTA by identifying key terms matching that we, hereafter, call *source words*. A case study in the food and beverage industry is provided to demonstrate the algorithm's application. In the study, five relevant knowledge areas were defined to guide literature selection. The database from scientific repositories was categorized using six classification rules based on impact factor (IF), Open Access (OA) status, and JCR journal ranking. This classification revealed the knowledge area with the highest presence and highlighted the effectiveness of the selection rules in identifying articles for the SOTA. The approach included a panel of experts who confirmed the algorithm's effectiveness in identifying source words in high-quality articles. The algorithm's performance was evaluated using the F_1 Score, which reached 0.83 after filtering out non-relevant articles. This result validates the algorithm's ability to extract significant *source words* and demonstrates its usefulness in building the SOTA by focusing on the most scientifically impactful articles.

Keywords: state-of-the-art; machine learning; knowledge management; keyword extraction; systematic literature review

1. Introduction

The collection and storage of scientific information have undergone significant advancements with the evolution of technology. However, the exponential growth of scientific

production has gradually rendered some traditional methods inefficient (Morales-Alarcón et al., 2024). Each year, over 8.8 million experts worldwide produce more than 2.5 million new research publications, reflecting an impressive annual growth rate of 8% to 9% (Morales-Alarcón et al., 2024). Broadly, there are two primary methods for disseminating knowledge in published works: depositing them in subscription-based repositories or making them available as open research in Open Access (OA) repositories. The latter approach is increasingly dominating knowledge dissemination, as major publishing companies are shifting from traditional subscription-based models to OA or hybrid models (Turgel & Chernova, 2024). This openness is quickly becoming the norm, supported by many leading universities transitioning to an open development model, steadily increasing their OA publications and prioritizing the transition (Turgel & Chernova, 2024). Moreover, a significant benefit of publishing in OA is that these articles tend to receive more citations than their subscription-based counterparts, while authors also experience reduced processing times and quicker receipt of initial citations (Abdin & De Pretis, 2024). Furthermore, developed countries like Japan have already announced that all publicly funded research will be made freely available (Ikarashi, 2023). This trend is one of the primary reasons why our research focuses on OA repositories while still considering subscription-based journals. There is a wide variety of OA repositories, including arXiv, bioRxiv, chemRxiv, medRxiv, SSRN, PsyArXiv, HAL, Preprints.org, EarthArXiv, OSF, and EpiPreprints. Many of these repositories are accessible through an application programming interface (API) that facilitates data extraction and automation, as exemplified by the arXiv API (arXiv API Access, 2025), which is incorporated into our framework.

Recent studies have demonstrated that AI can accelerate systematic literature reviews, enhance article selection, and reduce the manual workload involved in indexing (de la Torre-López et al., 2023). Accurate indexing is essential for ensuring that information is categorized and easily retrievable. It requires consistent terminology; a clear hierarchical structure; and, when necessary, the use of automated tools to manage large datasets efficiently (Manning et al., 2008; Sarthi et al., 2024). Effective indexing demands careful attention to elements such as titles, abstracts, summaries, and tables of contents, as outlined in the ISO 5963 (International Organization for Standardization, 1985) and UNE 50-121-91 standards, the latter specifically applying to Spanish-language documents (Asociación Española de Normalización y Certificación, 1991).

Models based on natural language processing (NLP) and deep learning have led to significant improvements in the organization of scientific information, facilitating easier access and classification (Musslick et al., 2025). Although automation has optimized several stages of the process, challenges persist in terms of model interpretability and the need for human oversight to ensure the quality of the processed information.

The evaluation of information retrieval systems dates back to Cyril Cleverdon's development of precision and recall metrics (Cleverdon, 1984). This measurement is particularly important given the inherent subjectivity, as the overlap between terms assigned by different individuals can vary between 30% and 60% (Cleverdon, 1984). Cleverdon's work, which quantified the proportion of relevant documents among the retrieved documents, became a foundational contribution in the 1960s (Jaiswal et al., 2024).

Knowledge management continues to be a critical aspect of the scientific community, enabling the efficient acquisition, construction, and communication of information (Zamiri & Esmaeili, 2024). In this context, the development of automated tools for collecting and analyzing scientific literature represents a strategic advance that can improve the effectiveness and efficiency of knowledge production.

A systematic review aims to identify and describe the current state of the art (SOTA) and AI techniques employed to represent and infer knowledge, enabling the efficient

manipulation of text and the ability to learn from large datasets (de la Torre-López et al., 2023). These techniques are particularly beneficial for tasks that are labor-intensive or repetitive for humans, such as analyzing scientific literature. Manually preparing and writing a systematic literature review (SLR) requires substantial time and effort, as it involves strategic planning, the performance of literature searches and analyses, and the reporting of findings (de la Torre-López et al., 2023). Additionally, scientific cartography, or science mapping, offers valuable tools for visualizing and analyzing relationships, trends, and structures within scientific disciplines. Tools like VOSviewer (Kirby, 2023) allow for graphical representations of geographic distribution, researcher collaborations, areas of interest, and the evolution of research topics.

A recent and highly significant development in the field of NLP is the emergence of large language models (LLMs) such as ChatGPT-4.5, Claude 3.5, Gemini 1.5, DeepSeek-R1, Meta-Llama 2, which have dramatically transformed the landscape of NLP (Alkhamissi et al., 2023; Miao & Holmes, 2023). The literature highlights several prompt engineering techniques applied to LLMs for academic document analysis, including the use of structured and context-specific prompts, each tailored for particular information-processing tasks (Liu et al., 2023). However, studies have documented a decrease in efficiency when LLMs are tasked with more complex assignments (Alshami et al., 2023; Schick & Schütze, 2021). Furthermore, the accuracy of these models is highly dependent on the design of task-specific prompts (Vatsal & Dubey, 2024). Different tasks demand varying levels of information and constraints to guide the LLM in producing the intended output (Sivarajkumar et al., 2024). As such, it is essential to design precise, relevant prompts rather than relying on generic or vague instructions, which can lead to confusion or erroneous results (Alshami et al., 2023; Sivarajkumar et al., 2024). Another challenge is inherent in the very nature of large-scale models like GPT-3, which, although capable of learning effectively from few examples, require substantial computational resources for training and operation (Wu et al., 2021). A notable comparison can also be drawn with the work of Ofori-Boateng et al. (2024), which introduced a fully automated method for keyword extraction. While this approach helps identify relevant terms, our research builds upon it by embedding keywords within the specific context of the academic field, thereby improving the accuracy of document selection. The performance of our system highlights its ability to retrieve a large volume of relevant documents with high accuracy, comparing favorably with existing approaches in the literature. However, achieving a balance between accuracy and completeness remains a persistent challenge.

1.1. State-of-the-Art Selection Using Artificial Intelligence

Panayi et al. (2023) assessed the use of advanced NLP models such as BERT and BiLSTM for the automated extraction of data from scientific publications related to systematic literature reviews (SLRs) in oncology and Fabry disease. The methodology is illustrated in Figure 1. Manual annotations of keyword were used to train the models, which were then compared with human annotations to assess their accuracy. Additionally, BERT was pre-trained with biomedical literature, and entity classification was optimized using a CRF model, with the F_1 Score metric employed to evaluate performance. The results indicated that the pre-trained BERT+CRF model performed best in entity recognition, achieving an accuracy of 73% in oncology and 70% in Fabry disease (Panayi et al., 2023). The model showed higher accuracy in identifying key clinical metrics, although it had lower precision in extracting treatment doses. On the other hand, the relationship classification model based on BERT achieved over 90% accuracy in cases with sufficient examples of relationships between biomedical entities (Panayi et al., 2023). These findings suggest that, with further adjustments, machine learning could become a key tool for the automated extraction of

data in systematic reviews, significantly optimizing processing times and reducing the manual workload involved in analyzing large volumes of scientific literature.
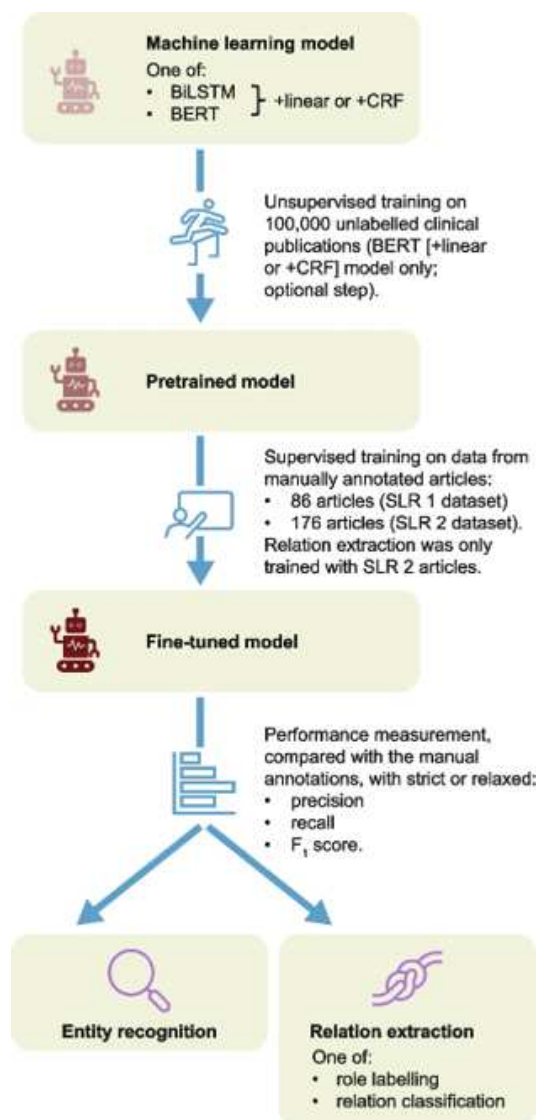

Figure 1. Process for enhancing entity recognition and relationship extraction in language models. Source: Panayi et al. (2023).

Despite advances in the automation of the building of cutting-edge technology, traditional methods of searching for and classifying information still require a considerable investment of time (Joseph & Ravana, 2024). Traditional methods like pooling, sampling, and using evaluation metrics for large document collection are time-consuming and have higher computational costs (Joseph & Ravana, 2024). Recent research has shown that integrating artificial intelligence into these processes can significantly improve efficiency (Golinelli et al., 2022). The COKE (COVID-19 Knowledge Extraction) project showcases the power of artificial intelligence in enhancing the systematic review of scientific literature (Golinelli et al., 2022). By leveraging machine reading models, deep learning, and sentence classification, the system effectively organizes large volumes of information into a structured format, enabling quicker and more informed decision making in high-uncertainty situations like the COVID-19 pandemic (Golinelli et al., 2022).

The implementation of advanced NLP and semantic clustering techniques has significantly improved the accuracy of identifying relevant information within large scientific datasets (Golinelli et al., 2022). Likewise, it has been discussed how the automation of

scientific practice, through machine learning and deep learning tools, has made it possible to optimize knowledge management and improve the reproducibility of scientific studies (Musslick et al., 2025). However, the need for human supervision in these processes is also noted, since data interpretation remains a critical factor for decision making in research (Musslick et al., 2025).

1.2. Selecting the State of the Art (SOTA) of a Patent Using a Textual Similarity Search

The method shown in Figure 2 focuses on the automated search for and selection of patents. It provides a detailed analysis of the strategies applied over the past decade for prior art retrieval in patents, with particular emphasis on advancements in NLP techniques and semantic search (Ali et al., 2024). Moreover, it addresses the main challenges affecting the accuracy and coverage of these searches and discusses potential future directions to enhance the effectiveness of information retrieval systems in this domain (Ali et al., 2024).

Systematic literature reviews have proven to be key tools for analyzing recent advancements and trends in patent information retrieval (Ali et al., 2024). In this context, various types of searches are necessary throughout the life cycle of a patent, from its conceptualization to its post-grant management. Figure 2 illustrates the key stages where patent information retrieval is required, including the initial search for the SOTA, the evaluation of patentability during the pre-application and examination stages, and technological monitoring and portfolio assessment in the post-grant phase (Ali et al., 2024).
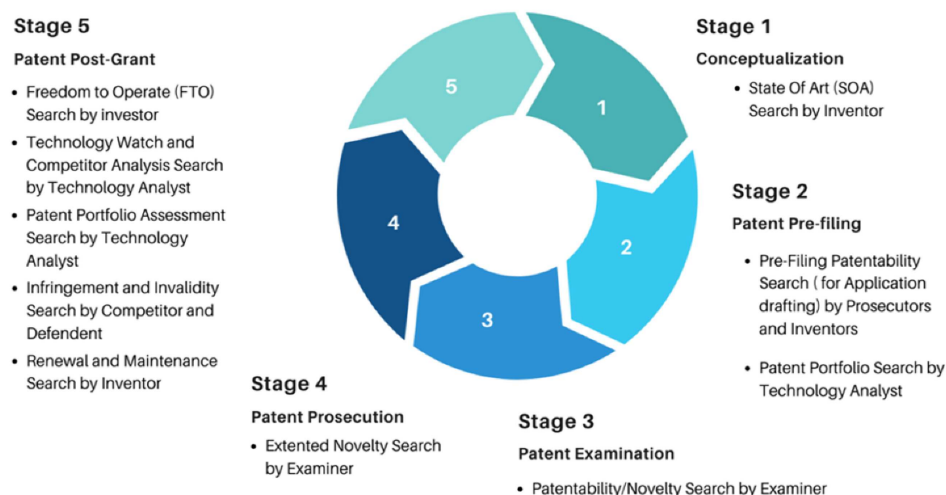

Figure 2. Patent research to identify related documents at various stages of their life cycles. Source: Ali et al. (2024).

Recent research underlines the need to apply advanced NLP and machine learning techniques for tasks such as article selection and full-text analysis, identifying challenges in transparency and replicability (Jiang et al., 2024). Likewise, in the same article, it was shown that extended language models can transform the collection and synthesis of information in scientific reviews, significantly reducing manual workload and improving accuracy in extracting key data (Jiang et al., 2024).

In this paper, a method that helps researchers quickly find relevant articles by searching for specific keywords in abstracts from scientific repositories is presented. To differentiate the term keywords, which is often used at the end of the article abstract, from keywords found in the abstract itself or in other metadata, such as the body of the article, in its title, etc., we refer to this term from now on as *source words*. The intention is to create a summary of articles based on the number of times these *source words* appear, making it easier to

identify important texts for research. The algorithm saves time by overcoming challenges like finding synonyms and handling subjective terms.

1.3. Aim and Scope

The aim of this research is to automate the process of gathering relevant articles to construct a comprehensive SOTA review in a specific field of knowledge. This is achieved by automating keyword searches within article abstracts, using an innovative algorithm that identifies the most relevant articles based on the frequency of *source words*. Unlike previous studies exploring the automation of systematic reviews in scientific literature through the use of artificial intelligence, our approach stands out by incorporating more efficient indexing via the generation of key terms and the systematic organization of documents.

The scope of the research is extensive, with manuscripts sourced from both subscription-based scientific repositories, such as Scopus, Web of Science, and IEEE Xplore, and open-access repositories accessed via API (e.g., arXiv). Additionally, specific search criteria, including impact factor (IF), open access (OA) status, and JCR journal classification, are incorporated into the approach. These criteria are considered indicative but not exhaustive.

1.4. Framework

The methodological steps involved in the proposed framework to find and choose pertinent articles to build the SOTA are shown in Figure 3. Defining the research problem and knowledge structure constitutes the initial step of our framework. This step establishes the criteria for selecting the *source words*, which are essential components for conducting an organized bibliographic search.
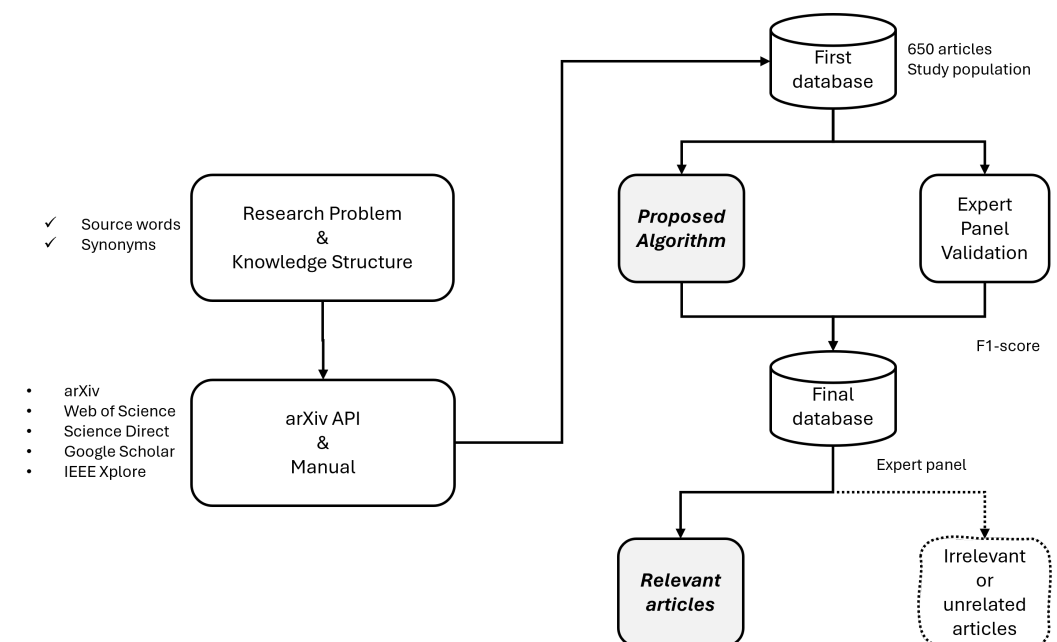
Figure 3. Proposed framework (created by the authors).

The selection of *source words* should prioritize terms that capture the core semantics of the topic while avoiding ambiguous or overly generic words. The identification of *source words* and their synonyms includes (1) defining the problem and its scope, (2) identifying the key concepts of the problem, (3) selecting the main areas of knowledge and their subtopics, and (4) generating *source words* and synonyms for each subtopic.

The next step is the search for and gathering of scientific data, which is done in two complementary ways: automatically by using an Application Programming Interface (API)

in the arXiv repository (arXiv API Access, 2025) and manually by consulting a variety of specialized academic repositories. This two-pronged strategy ensures thorough and organized coverage of the study's pertinent literature. The first research database, also known as the study population is then created, consisting, in our case, of 650 articles in total. The algorithm presented in his article is applied to the articles that comprise the initial database of the study to determine the documents that will comprise the study's final database; simultaneously, a panel of experts conducts a validation process that is applied to a statistical sample that is chosen at random with a 90% confidence level. The F_1 Score metric, which is a measure of precision and recall in the analysis of outcomes, is determined as a result of this validation. The procedure ends with the selection of papers deemed relevant, which serve as a crucial foundation for the development of the SOTA. In turn, papers that are deemed irrelevant or unrelated to the study's goals are identified and discarded by the expert panel.

The proposed framework helps researchers in the selection of literature for the SOTA. The proposed algorithm is applied until the researcher has a collection of information that he considers important for his research, only by searching in scientific repositories, without reading each article in depth. A technical sheet is then prepared in table format, usually in Excel, which includes columns for the title and abstract of each article. The algorithm performs specific *source word* searches in the abstracts listed in the table, recording the rows in which such words are found. This procedure is repeated with all the defined *source words*, and the algorithm generates a summary with the *source words* found in each article. The texts selected as part of the SOTA are those with the highest number of *source-word* matches, allowing the user to decide the selection threshold.

The proposed algorithm significantly facilitates the identification of the SOTA by addressing challenges such as identifying synonyms or abstract terms related to the *source words*, thee subjectivity of terms assigned by different people, and the time needed to identify relevant information. In our case of study, the algorithm was applied to develop a system that assesses wine quality based on the analysis of color. Based on the research objectives and expert opinion, five key areas were defined: structural equations, sensory analysis, food and beverages, colorimetry, and artificial intelligence. For each of these areas, *source words* were established that the algorithm used to search in the abstracts of the selected articles. The employed methodology is further explained in Section 2.

This manuscript is organized as follows. Section 1 provides a review of the SOTA and a presentation of our framework. Section 2 presents the methodology of the framework, whose results of application to a case study are presented in Section 3. Finally, the conclusions, limitations, and future work are described in Section 5.

2. Methodology

One of the main approaches leading to a framework for research on a topic or field is content analysis, which can help reduce systematic error and random effects and provide more reliable results according to the method proposed by Hachicha and Ghorbel for the formation of frameworks, which consists of two phases (Hachicha & Ghorbel, 2012). The first phase is the definition of sources for the search for articles, and the second is the classification of the selected articles. For our research, this method is used, and a third phase is added, which includes classification using fuzzy rules.

2.1. First Phase: Bibliographic Search

1. The main themes that would determine the bibliographic search were determined by the titles of the articles. The topics that were identified were sensory analysis, physical–chemical analysis, wine evaluation, and structural equations. In relation to

the identified topics, a set of *source words* was established for the bibliographic search: color spaces, colorimetry in food, wine, wine color, wine quality, artificial vision, sensory analysis, structural equations, fuzzy logic, artificial intelligence, sensory quality, color spectrum, CIELab, and color measurements.

2. The computerized search of the set of *source words* was limited to scientific articles from the literature and scientific repositories, such as ScienceDirect, WoS, Goggle Scholar, Springer Link, and IEEE Xplore. In addition, using the same *source words*, the search also included articles extracted from the arXiV repository using an available API. Through a computerized search with the specifications and in the scientific repositories mentioned above, a population of 650 articles was obtained.

3. The classification rules for the population of articles were established using two evaluation criteria (as shown in Table 1).

Table 1. Evaluation criteria for the population of items.

Criterion	Outputs
OA	Yes, No
IF	<1, 1 & 2, >2

The OA criterion is represented by two possible outputs: "Yes" for articles that are open-access and "No" for articles that have restricted access (by subscription). The IF (impact factor) criterion is represented by three outputs: "IF < 1" for articles published in journals with an impact factor of less than 1, "IF 1 and 2" for articles published in journals with an impact factor between 1 and 2 and, and "IF > 2" for articles published in journals with an impact factor greater than 2. These criteria were considered fundamental for the selection of scientific articles. Table 2 presents the six resulting rules.

Table 2. Classification rules that are effective for research.

Rule Number	Classification Rules
R1	(I.F. < 1) ∩ (O.A. = Yes)
R2	(I.F. < 1) ∩ (O.A. = No)
R3	(I.F. 1 and 2) ∩ (O.A. = Yes)
R4	(I.F. > 2) ∩ (O.A. = Yes)
R5	(I.F. 1 and 2) ∩ (O.A. = No)
R6	(I.F. > 2) ∩ (O.A. = No)

Figure 4 shows the behavior of the population of articles when applying the six classification rules. In this figure, it can be observed that category R1 represents the largest proportion of articles, with a value of 400, followed by R5 with 168 and R6 with 72.

In contrast, categories R3 and R4 have a significantly lower share, with values of 2 and 8, respectively. It is worth mentioning that R2 was omitted in this graph because it did not contain any articles. The different colors allow each category to be identified in the figure, which facilitates the interpretation of the distribution. It is important to note that the application of the classification rules (R1–R6) may is optional for each project; however, in this case it was decided to use them to maintain better order in the articles that belong to each of the classification rules, allowing for a better organization and analysis of the data.
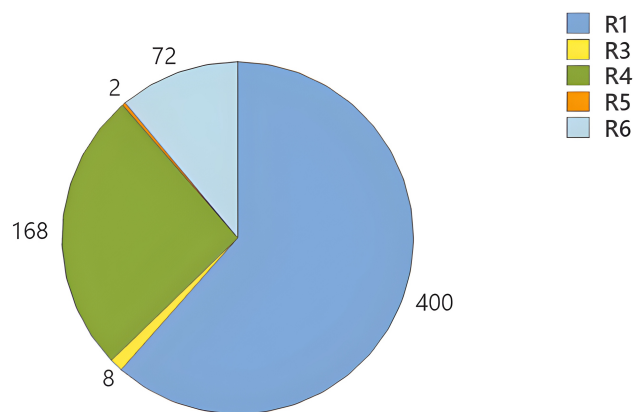
Figure 4. Behavior of the article population applying the classification rules.

2.2. Second Phase: Literature Classification

Based on the titles of the articles, the areas of knowledge relevant to the creation of the reference framework were determined, as well as their respective codes, as shown in Table 3. After carrying out the first phase, the need to assign weights to the articles included in each rule was evaluated. The purpose of this was to evaluate the frequency with which the *source words* (Table 4) appear in the areas of knowledge relevant to the study. The identification of the *source words* was achieved through the proposed algorithm developed in Python 3.12.

The operation of this algorithm is provided in Algorithm 1.

Algorithm 1 Search for *source words* in a document

1: **Define:** Title of the document and column where summary is located
2: **Define:** Sheet name to be read and the source word to be searched for
3: **Input:** Sheet name, column location, source word
4: **Search:** Search for the source word in the corresponding column
5: **Store:** Store the row number where the source word is found
6: **Output:** Display the numbers of all rows where the source word is found

---

Table 3. Areas of knowledge relevant to research.

Code	Area
A	Structural equations
B	Sensory analysis
C	Food and beverages
D	Colorimetry
E	AI, pattern recognition, and artificial vision

Table 4. *Source words* by area of knowledge.

Code	Source Words
A	Structural equations and SEM
B	CATA, RATA, smell, sensory, color, and flavor
C	Food, beverages, wine, and red wine
D	Colorimetry, sprectrophotometer, colorimeter, CIELab, color, and colorimetric
E	Pattern recognition, artificial vision, AI, Automated learning, artificial intelligence, and intelligent

2.3. Third Phase: Classification of Literature Using a Relationship Diagram

To classify the literature, it is necessary to create a relationship diagram in which the relationships between each area of knowledge can be observed. Keep in mind that the knowledge areas were chosen by subject-matter experts who also established the connections between them based on how crucial it is to include a combination of each area in an article. In Figure 5, each relationship is represented by a line; areas with a high degree of relationship are joined with four lines, those with a medium degree of relationship are with three lines, and those with a low degree of relationship are joined with two lines.
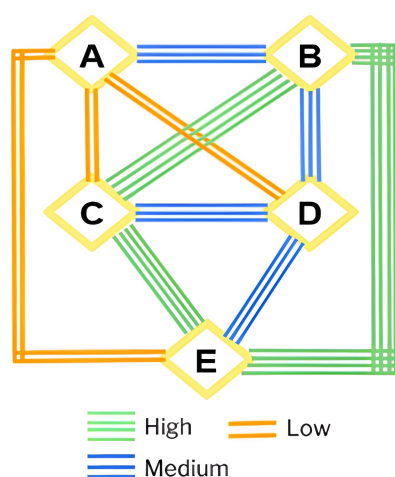

Figure 5. Diagram of relationships between the areas of knowledge of the topic to be investigated.

For example, if an article contains a relationship between areas of knowledge B and C, it is considered to exhibit a high degree of relationship; if it contains a relationship between areas of knowledge D and E, it is considered to have a medium degree of relationship; and if it contains a relationship between areas of knowledge A and D, it is considered a to have low degree of relationship.Each line that represents a relationship between areas of knowledge has a value of seven points. If the relationship is less relevant, one or two points are deducted, depending on the level of importance.

Based on the above considerations, each relationship is assigned a weight ranging from 13 to 28 points. This assignment reflects the degree of interest (Z) of the article for research, where 13 represents the lowest level of interest and 28 is the highest (Table 5). These weights were defined by experts in the field, classifying relationships with values between 13 and 14 as of low interest, those between 19 and 21 as of medium interest, and the rest as of high interest (Table 6).

The inference rules are determined according to the degree of interest to which each relation of the areas of knowledge belongs (Figure 6). These inference rules are based on the Mamdani method, which indicates that logical propositions comply with the "if–then" rule. For this research, three rules were obtained, which are described below:

- If a scientific article contains a relationship between B and C, B and E, or C and E, it is considered to have a high degree of interest (Z) for this research.
- If a scientific article contains a relationship between A and B, B and D, C and D, or D and E, it is considered to have a medium degree of interest (Z) for this research.
- If a scientific article contains a relationship between A and C, A and E, or A and D, it is considered to have a low degree of interest (Z) for this research.

Table 5. Correlation of each area of knowledge relevant to the research.

	A	B	C	D	E
Fuzzy Sets in Correlation Table					
A		19	14	13	14
B			28	21	28
C				20	27
D					20
E					

Table 6. Degree of interest according to the weighting of the relationships between the research areas.

Points	Z
Linguistic Values for Measuring Degree of Interest with Z Research	
13–14	Low
19–21	Medium
27–28	High

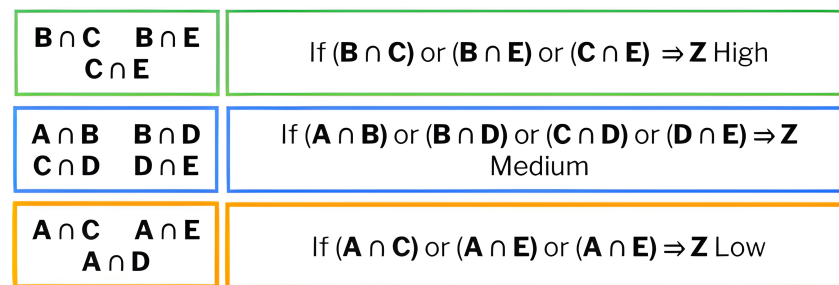| **B ∩ C B ∩ E** **C ∩ E**	If (**B ∩ C**) or (**B ∩ E**) or (**C ∩ E**) ⇒ **Z** High
A ∩ B B ∩ D **C ∩ D D ∩ E**	If (**A ∩ B**) or (**B ∩ D**) or (**C ∩ D**) or (**D ∩ E**) ⇒ **Z** Medium
A ∩ C A ∩ E **A ∩ D**	If (**A ∩ C**) or (**A ∩ E**) or (**A ∩ E**) ⇒ **Z** Low

Figure 6. Mamdani's inference rules for determining the degree of interest of each article.

3. Results

After searching for *source words* using the Algorithm 1, they were related to each classification rule as shown in Figure 7, where the distribution of the number of scientific articles according to the classification rules (R1 to R6) and areas of knowledge (A, B, C, D, and E) is shown.
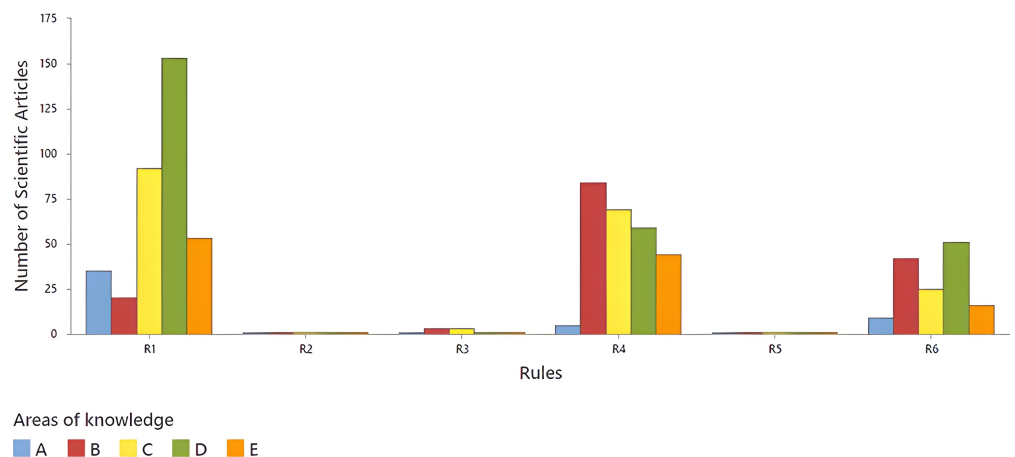

Figure 7. Analysis of the literature by areas of knowledge and classification rules.

Each area of knowledge is represented by a color: structural equations (A) in blue, sensory analysis (B) in red, food and beverages (C) in yellow, colorimetry (d) in green, and artificial intelligence (E) in orange.

For an article to be relevant to this research, it must cover three knowledge areas. It is important to note that in some cases, several *source words* belonging to the same area were found in the same article, suggesting a closer relationship with the research topic. It can be observed that rules R1 and R5 presented the highest number of articles, while R2, R3, and R4 presented very low values. In R1, area of knowledge "D" (green) dominates, with the highest number of articles, followed by "C" (yellow) and "E" (orange). These data indicate that certain rules concentrate scientific production in specific areas, while others present a lower number of articles in all categories.

The frequency of the appearance of *source words* in each knowledge area was analyzed. The analysis was performed for each classification rule, starting with the evaluation of the articles contained in R1 (Figures 8–12 are presented in this paper as examples).

Figures 8–12 show stacked bars displaying the repetition of *source words* in different scientific articles, identified by their row number on the X-axis. Each bar represents an article and is segmented into five knowledge areas (A, B, C, D, and E), differentiated by colors. It can be observed that the frequency of *source-word* repetition varies between articles, with some reaching up to four repetitions, while others have lower values. Knowledge areas D and E seem to be predominant in most articles, while C and B have lower levels of representation. Furthermore, it can be noticed that, in certain rows, such as 34 and 18, there is a higher concentration of *source words* compared to others. This analysis allows for the identification of the distribution of *source words* within articles according to the different knowledge areas, providing a structured view of the thematic classification in this dataset. The first article to highlight in Figure 8 is located in row 2, since it covers knowledge areas B, C, and D. The next articles to highlight are those in rows 14, 15, and 18, which cover areas B, D, and E.
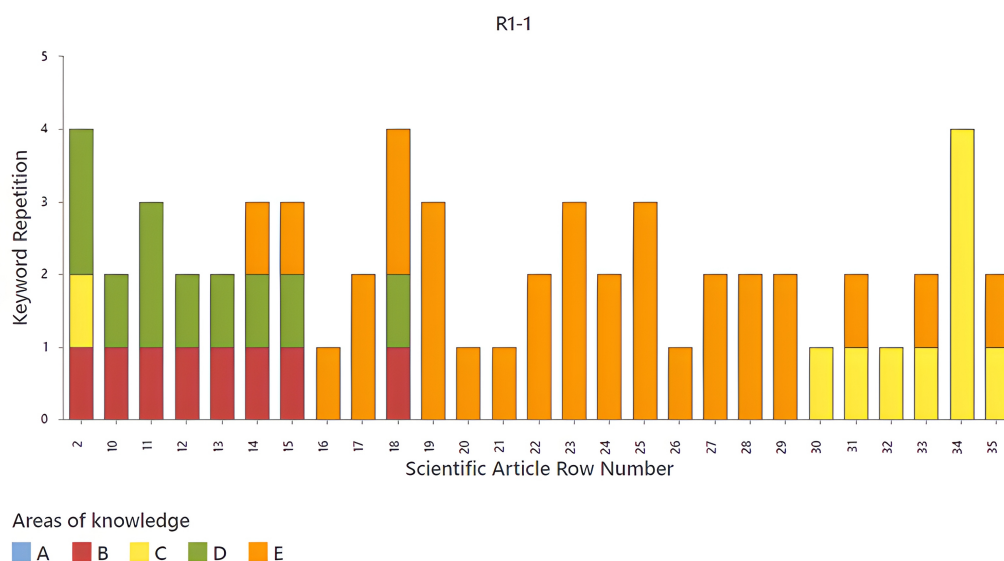

Figure 8. Descriptive analysis of the repetitiveness of the *source words* in each article in R1 (R1-1).

In Figures 9 and 10, it can be observed that most of the articles present combinations of two knowledge areas. In Figure 9, the most recurrent are C and E in the first half of the graph and B and D in the second half, and the opposite occurs for Figure 10. The articles corresponding to rows 57, 81, 125, and 240 stand out, since they contain three knowledge areas simultaneously, suggesting a greater interdisciplinarity in these studies. The variability in the repetition of *source words* and the combination of areas allow for the identification of trends in the thematic classification of the analyzed articles. In Figure 11, it can be observed that most of the articles present combinations of three areas of knowledge (B, C, and E). The articles corresponding to rows 272 and 302 stand out, since they contain

three areas of knowledge simultaneously, suggesting a greater interdisciplinarity in these studies. In Figure 12, it can be observed that most of the articles are represented by area of knowledge A. However, in this dataset, no article stands out, since none presents three areas of knowledge.
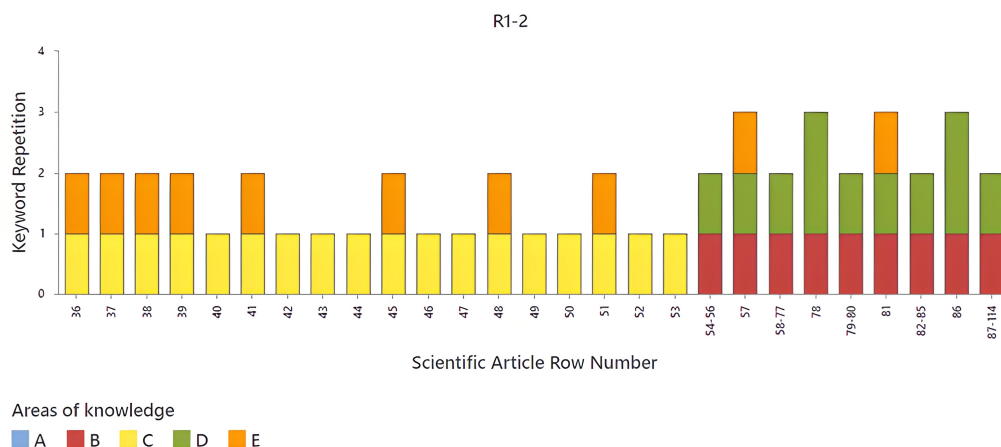

Figure 9. Descriptive analysis of the repetitiveness of *source words* in each article in R1 (R1-2).
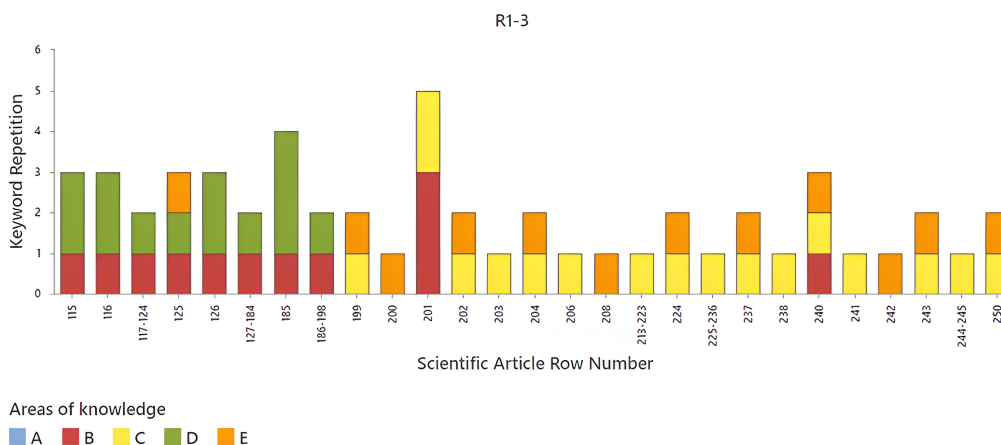

Figure 10. Descriptive analysis of the repetitiveness of *source words* in each article in R1 (R1-3).
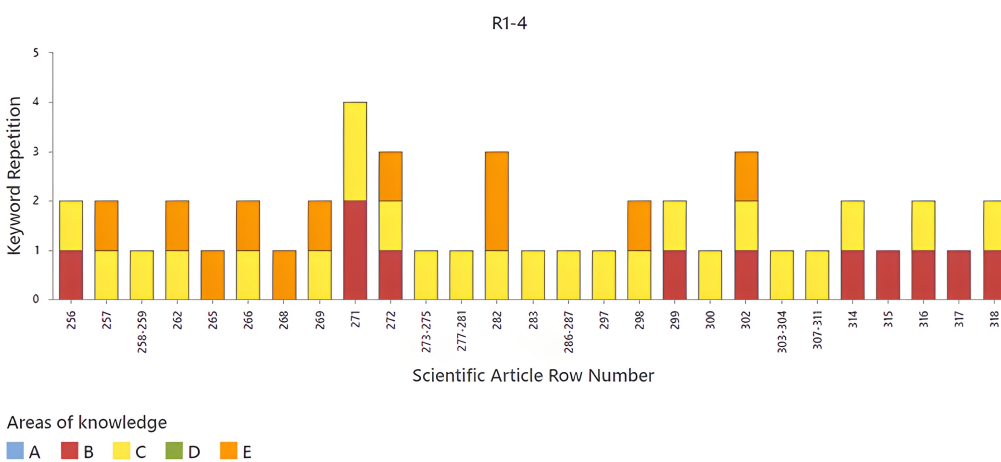

Figure 11. Descriptive analysis of the repetitiveness of *source words* in each article in R1 (R1-4).
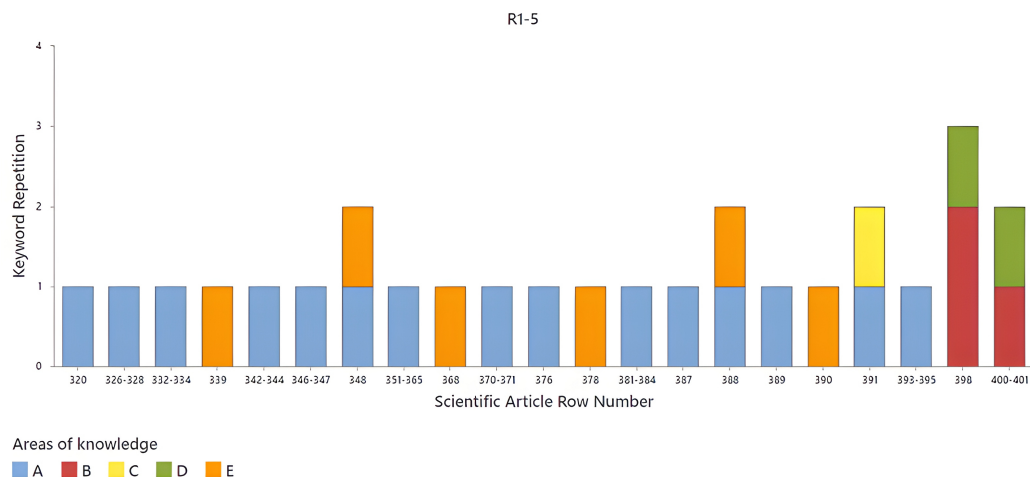
Figure 12. Descriptive analysis of the repetitiveness of *source words* in each article in R1 (R1-5).

The second rule lacks information, as no article is found, so the next to be analyzed is R3 (Figure 13). The content of articles for this rule is small, with only six in total. Of these, the following stand out: the article corresponding to row three, which presents *source words* in areas B and D, and the articles in rows 5 and 6, which cover areas B and C, with a *source words* in each. It is important to note that the most prominent knowledge area represented in this rule is sensory analysis (B). However, since these articles do not cover three knowledge areas, they are not considered important for this research.
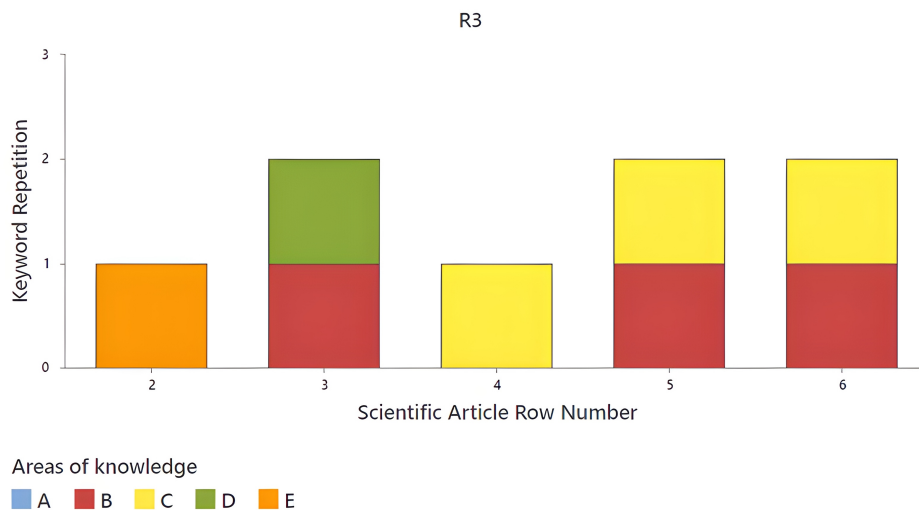

Figure 13. Descriptive analysis of the repetitiveness of the *source words* in each article in R3.

The next rule to be analyzed is R4. Most of the literature is found in this rule, so it is necessary to represent the analysis in three figures. The first set observed in Figure 14 comprises articles presented between row 1 and row 61 of the Excel sheet. It is worth mentioning that the articles where no *source words* was found are omitted. Among this set of articles, the one in row 39 stands out, with *source words* in each of the three areas of knowledge—B, D, and E. Therefore, it is evident that in this set, the areas of knowledge are sensory analysis (B), colorimetry (D), and artificial intelligence (E).
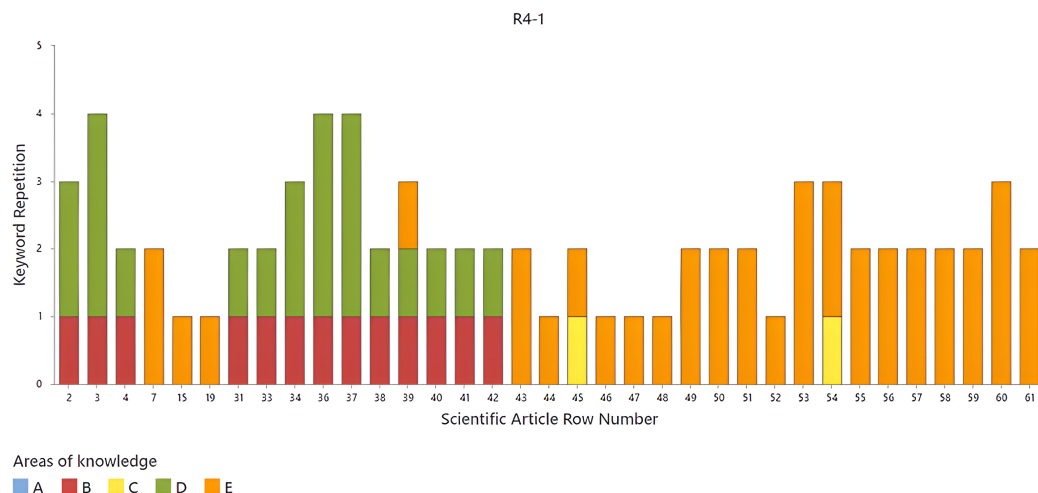
Figure 14. Descriptive analysis of the repetitiveness of *source words* in each article in R4 (first set).

Continuing with the analysis of R4, Figure 15 shows the second set of articles. This set includes the articles from row 62 to row 98. Likewise, the articles in which no *source words* were found in the abstract are omitted. In this set, several articles stand out that encompass three areas of knowledge (B, C, and D). These articles are found in rows 66, 67, 68, 74, 75, 76, 77, 84, 90, 91, 95, 96, and 98. From this set, a total of thirteen important articles for the research were obtained, covering the areas of knowledge of sensory analysis (B), food and beverages (C), and colorimetry (D).
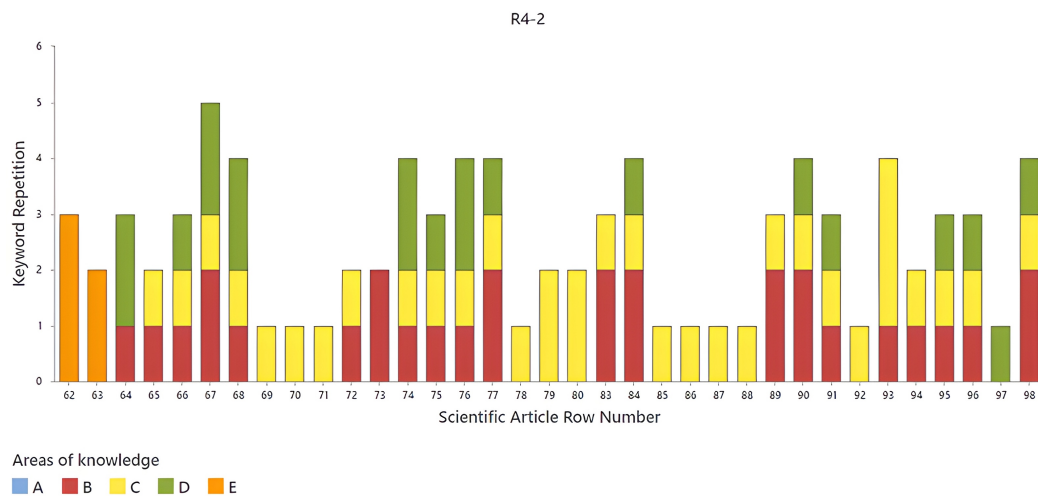

Figure 15. Descriptive analysis of the repetitiveness of *source words* in each R4 article (second set).

To conclude the analysis of R4, Figure 16 shows the third cluster of articles. This cluster is represented by the articles from rows 99 to 135, omitting the articles where no *source words* were found. It can be observed that most of the articles contained in this cluster cover three areas of knowledge. These areas of knowledge are the same as the for the previous cluster (B, C, and D). In addition, in this set of articles, there is one in row (134) that does not contain area of knowledge D; however, it does contain area of knowledge A, which corresponds to structural equations. This shows that the main topics are structural equations (A), sensory analysis (B), food and beverages (C), and colorimetry (D).
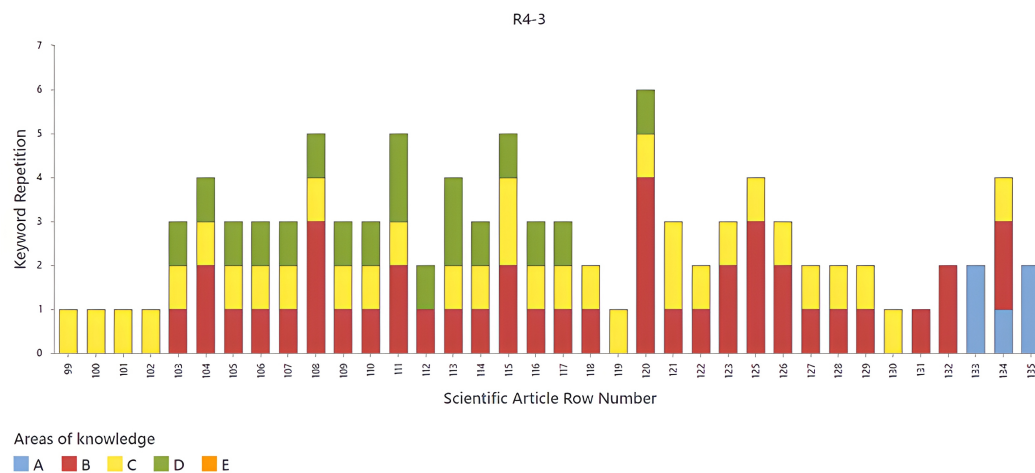
Figure 16. Descriptive analysis of the repetitiveness of *source words* in each article of R4 (third set).

In the following rule (R5), there is only one article (Figure 17), and the identified *source words* belong to the areas of colorimetry and artificial intelligence (D and E, respectively). This article only covers two areas of knowledge; therefore, it is not considered important for the research.
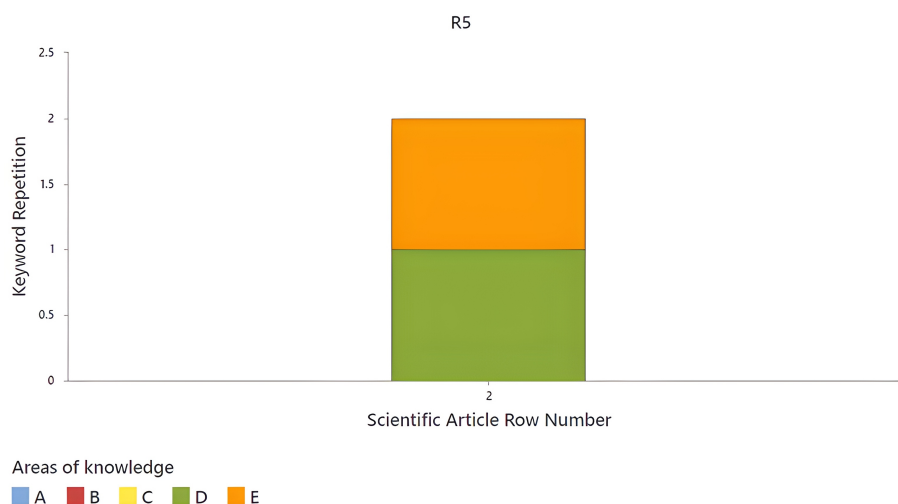

Figure 17. Descriptive analysis of the repetitiveness of the *source words* in each article of R5.

Finally, the analysis of R6 is presented. This analysis is divided into two sets of articles (Figures 18 and 19), the first of which contains the articles from rows 2 to 32 and the second of which contains the articles from rows 33 to 61. In both sets, the articles in which no *source words* were found were omitted. The articles with the greatest relevance for this research that belong to the first set (Figure 18) are the articles corresponding to rows 2, 12, 16, 17, and 20. These articles cover three areas of knowledge; the article in row 12 covers areas B, D, and E, and the remaining articles ones cover areas B, C, and D.

The second set of R6 articles is shown in Figure 19. It can be observed that the articles to be highlighted are those in rows 44, 45, 46, 49, 51, 53, and 55, covering the same three areas of knowledge: B, C, and D.
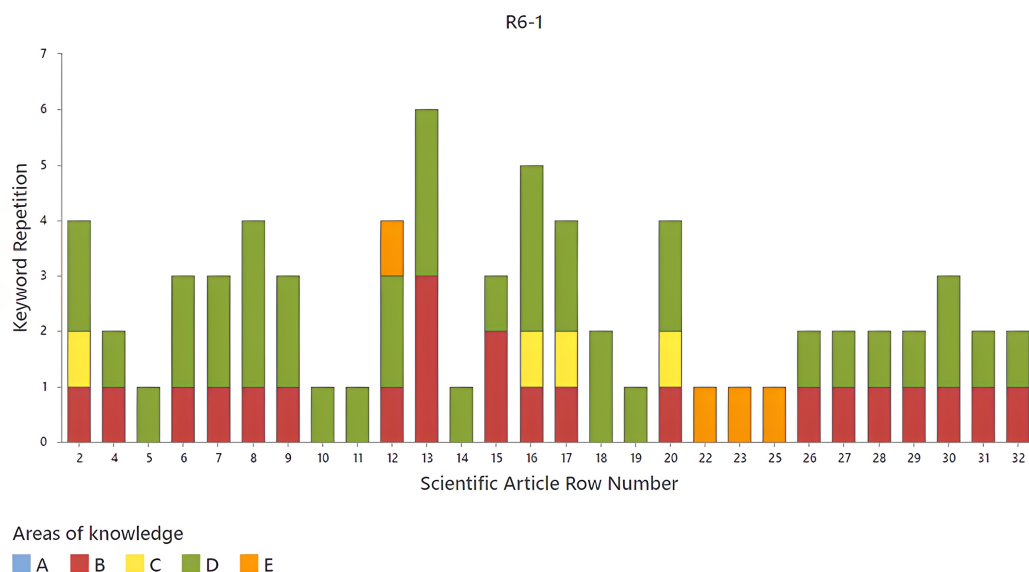
Figure 18. Descriptive analysis of the repetitiveness of the *source words* in each article of R6 (first set).
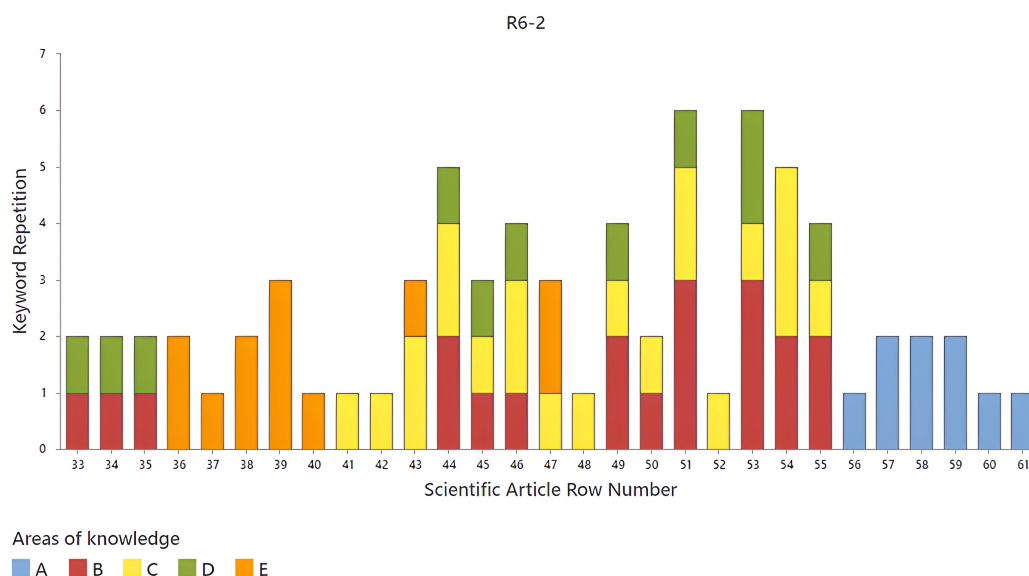

Figure 19. Descriptive analysis of the repetitiveness of the *source words* in each R6 article (second set).

Therefore, for R6, the areas of knowledge are sensory analysis (B), food and beverages (C), colorimetry (D), and artificial intelligence (E). Comparing the analysis of each rule with Figure 9, it is evident that, in both instances, the areas of sensory analysis (B), food and beverages (C), and colorimetry (D) are those that present the greatest repetition of *source words*. This indicates that the majority of the located literature is linked to these three areas in particular.

3.1. Accuracy of a Classification Model

The F_1 Score is a metric used to evaluate the accuracy of a classification model by combining two fundamental measures—precision and recall—into a single value. It is considered a balanced measure between these two metrics that is especially useful when the data in the classes are imbalanced, meaning there is a dominant class. The reason for using the harmonic mean instead of the arithmetic mean between precision and recall is that the harmonic mean gives more weight to lower values. This means that the F_1 Score is low if either of these two metrics is low, penalizing models that are imbalanced between

precision and recall. Mathematically, the F_1 Score is defined as the harmonic mean between precision and recall.

The formula is expressed as follows:

$$F_1 \text{ Score} = 2 \times \frac{P \times R}{P + R} \tag{1}$$

where

$$P = \text{precision}, \quad R = \text{recall}$$

In turn, we consider

$$TP = \text{true positive}, \quad FP = \text{false positive}, \quad FN = \text{false negative}$$

Precision measures how accurate a model's positive predictions are. It is defined as the number of true positives divided by the total number of positive predictions:

$$P = \frac{TP}{TP + FP} \tag{2}$$

Recall measures the ability of the model to identify all positive cases. It is calculated as the number of true positives divided by the total number of true positives and false negatives:

$$R = \frac{TP}{TP + FN} \tag{3}$$

The case study considers a database with a total of 650 articles to be classified into two classes: articles relevant to a specific research topic and articles not relevant to the specific research topic. A panel of specialists in the field of the research topic, consisting of researchers (professors and PhD students), was consulted to review the database and obtain a benchmark classification of the articles it contains. Table 7 provides an overview of the profiles of the expert panel, while Table 8 outlines their professional backgrounds, average age, and gender distribution.

Table 7. Expert panel profiles.

Expert Profiles	Panel Information	
	No. of Experts	Average Years of Expertise
Public-sector institutions	7	16.6
Private sector	1	10
Food technology	1	20
Statistical technologies	3	25
Innovation technologies	7	12.5
Artificial intelligence technologies	6	11

The process of reviewing 650 articles requires time and monetary resources, which are limited in research. Therefore, classical statistics were used to review a portion of the 650 articles, selected as a random sample with 90% confidence and a margin of error of 0.05, in order to infer the results to the entire database. Thus, based on the sample size calculation, 100 articles were reviewed, selected through simple random sampling. The statistical approach for calculating the sample size was binomial, since only two classes were considered with the parameters shown in Table 9.

Table 8. Professional background.

Professional Background	Panel Information	
	No. of Experts	**Additional Details**
Enology and/or sensory evaluation	3	
Academia and industry	8	
PhDs/average years of experience	6	12.5 years
Doctoral students/research progress	2	20%
Average age		41.7 years
Gender distribution		2 women and 6 men

Table 9. A: Number of articles considered as the study population. **B:** Level of confidence for the statistical calculation of the sample size. **C:** Expected proportion of the population that has the expected classification. **D:** Expected margin of error. **E:** Z value (theoretical value of the standardized normal distribution) corresponding to the value of (B). **F:** Sample size calculation. *n* is the sample size used in the research. **G:** Number of articles classified as "not relevant by the panel of specialists" based on the value of "n". **H:** Number of articles classified as "relevant by the panel of specialists" based on the value of "n". **I:** Number of articles that are inferred to be classified as "not relevant by the panel of specialists" based on the value of "A". **J:** Number of articles that are inferred to be classified as "relevant by the panel of specialists" based on the value of "A".

A	B	C	D	E	F	n	G	H	I	J
650	90%	0.077	5%	1.96	76	100	68 (68%)	32 (32%)	442	280

3.2. F_1 Score Calculation

3.2.1. Defining the Ranking of Relevant Articles as a Success

In this case, TP is the number of articles that the algorithm classifies as relevant when they really are, that is, the experts classified them as such; FP is be the number of articles that the algorithm classifies as relevant when the experts classified them as not relevant; and FN is the number of articles that the algorithm classified as not relevant when the experts determined that they were. For the case study, the truth table and the corresponding calculations are given in Table 10:

$$P = \frac{TP}{TP + FP} = \frac{7}{7 + 1} = 0.875$$

$$R = \frac{TP}{TP + FN} = \frac{7}{7 + 25} = 0.2188$$

$$F_1 \text{ Score} = 2 \cdot \frac{P \cdot R}{P + R} = 2 \cdot \frac{0.875 \cdot 0.2188}{0.875 + 0.2188} = 0.35$$

Table 10. Truth table for relevant articles.

Panel: Relevant Article		Yes	No
Algorithm: Relevant Article	**Yes**	$TP = 7$	$FP = 1$
	No	$FN = 25$	$TN = 67$

3.2.2. Defining the Classification on Non-Relevant Articles as a Success

In this case, TP is the number of articles that the algorithm classified as not relevant when they are truly not relevant, i.e., the experts classified them as such; FP is the number of articles that the algorithm classifies as not relevant when the experts classified them as relevant; and FN is the number of articles that the algorithm classified as relevant when

the experts determined that they were not, i.e., not relevant. For the case study, the truth table and the corresponding calculations are given in Table 11:

Table 11. Truth table for Non-relevant articles.

Panel: Non-Relevant Article		Yes	No
Algorithm: Non-relevant article	**Yes**	$TP = 67$	$FP = 25$
	No	$FN = 1$	$TN = 7$

$$P = \frac{TP}{TP + FP} = \frac{67}{67 + 25} = 0.7283$$

$$R = \frac{TP}{TP + FN} = \frac{67}{67 + 1} = 0.9853$$

$$F_1 \text{ Score} = 2 \cdot \frac{P \cdot R}{P + R} = 2 \cdot \frac{0.7283 \cdot 0.9853}{0.7283 + 0.9853} = 0.8375$$

The results indicate that when the selection of non-relevant articles is defined as "success", the F_1 Score is higher than when the selection of relevant articles is defined as "success". Based on these data and the fact that the panel selected more relevant articles than the algorithm, it can be concluded that one of the following three approaches will have to be implemented in future research: i. train the algorithm to make it more flexible, ii. perform a stricter review by the panel, or ii. both strategies simultaneously. Lastly, it can be said that the algorithm can be used to filter out the articles that a human reviewer would deem irrelevant during the early phases of research. Finally, Table 12 shows a summary of the articles that were important for this research.

Table 12. Research papers and knowledge areas.

Title	Year	Row	R1–R6	A, B, C, D, and E
Distillation-Enhanced Physical Adversarial Attacks…	2025	57	1	B, D, E
Prompt-Guided Image-Adaptive Neural Implicit Lookup…	2024	81	1	B, D, E
Two improved algorithms for sparse generalized canonical…	2023	318	1	B, C, E
Professional Certification Benchmark Dataset: The First…	2023	272	1	B, C, E
Evaluating AI Vocational Skills Through Professional…	2023	240	1	B, C, E
Assessing ultrapremium red wine quality using PLS-SEM…	2023	134	4	A, B, C
An Ensemble Machine Learning Approach for Screening...	2023	125	1	B, D, E
Characterization of the color parameters and monomeric…	2023	116	4	B, C, D
Impacts of B-cyclodextrin bead polymer (BBP) treatment…	2023	115	4	B, C, D
The key role of vineyard parcel in shaping flavonoid…	2023	114	4	B, C, D
Effect of the addition of soluble polysaccharides from red…	2023	104	4	B, C, D
Metal-induced color change in blackberry wine…	2023	103	4	B, C, D
Modeling and application of sensory evaluation of...	2023	98	4	B, C, D
Characterization of the color parameters and monomeric…	2023	96	4	B, C, D
Grape-derived pectic polysaccharides alter the tannin...	2023	95	4	B, C, D
Application of salicylic acid to cv. Muscat Hamburg…	2023	91	4	B, C, D
Investigation of dynamic changes in quality of small...	2023	90	4	B, C, D
Cultivar difference characterization of kiwifruit wines…	2023	84	4	B, C, D
Pulsed electric field processing of green tea-infused…	2023	77	4	B, C, D
Discrimination of Port wines by style and age...	2023	76	4	B, C, D
The key role of vineyard parcel in shaping flavonoid...	2023	75	4	B, C, D
Non-acylated and acylated anthocynins in red wines...	2023	46	6	B, C, D
Impact of skin contact time, oak and tannin addition…	2023	45	6	B, C, D
Sensory evaluation of wine aroma...	2023	44	6	B, C, D

Table 12. *Cont.*

Title	Year	Row	R1–R6	A, B, C, D & E
Combination of effective color information...	2023	39	4	B, D, E
Combination of colorimetry, inner filter effect-induced...	2023	20	6	B, C, D
The standardized design and application guidelines...	2023	18	1	B, D, E
Exploiting microdistillation and smartphone-based...	2023	17	6	B, C, D
Headspace single drop microextraction based visual...	2023	16	6	B, C, D
IAACS: Image Aesthetic Assessment Through...	2023	15	1	B, D, E
A reconfigurable integrated electronic tongue and its use...	2022	302	1	B, C, E
Quality improvement of jujube wine through mixed...	2022	120	4	B, C, D
Effect of mannoproteins from different oenological yeast...	2022	113	4	B, C, D
Evaluation of the color and aroma characteristics...	2022	111	4	B, C, D
Interactions of Vitis vinifera L. cv. Pinot Noir grape...	2022	110	4	B, C, D
Improving effects of three selected co-pigments...	2022	109	4	B, C, D
Micro-oxygenation level as a key to explain the variation...	2022	108	4	B, C, D
Evolution of red wine in oak barrels with different...	2022	107	4	B, C, D
Effect of mannoproteins from different oenological yeast...	2022	68	4	B, C, D
Evaluation of the color and aroma characteristics of...	2022	67	4	B, C, D
Interactions of Vitis vinifera L. cv. Pinot Noir grape...	2022	66	4	B, C, D
Tongue diagnosis based on hue-saturation value color...	2022	14	1	B, D, E
Determination of peroxide value in edible oils based on...	2022	12	6	B, D, E
Anthocyanin profiles and color properties of red wines...	2021	106	4	B, C, D
Focus on the role of seed tannins and pectolytic enzymes...	2021	105	4	B, C, D
Representational and sensory cues as drivers of...	2021	51	6	B, C, D
Comparison of computer vision system and colour...	2021	2	1	B, C, D
Perception of quality and complexity in wine and their...	2020	55	6	B, C, D
Objective measures of grape quality: from Cabernet...	2020	49	6	B, C, D
Understanding quality judgments of red wines by...	2016	53	6	B, C, D
Adding oenological tannin vs. overripe grapes...	2014	2	6	B, C, D

4. Discussion

In the case study, five knowledge areas with research-relevant *source words* were identified for literature selection. The database was sourced from subscription-based scientific repositories and the open-access arXiv repository via an API (arXiv API Access, 2025). The search was structured using six classification rules (R1–R6) based on three criteria: impact factor (IF), open access (OA), and JCR journal ranking. The most represented knowledge areas included colorimetry, food and beverages, and sensory Analysis. Some classification rules were found to have fewer articles, indicating that certain rules are more influential or easier to meet. The repetition of *source words* played a key role in prioritizing articles, suggesting areas of greater research interest.

The figures representing the analysis of the articles in each classification rule demonstrate the thematic diversity of the selected articles, suggesting a multidisciplinary approach. These results confirm the effectiveness of the applied selection rules in identifying articles that contribute meaningfully to the SOTA. Out of 650 articles retrieved from the computerized search, 51 were deemed relevant for SOTA development (Table 12). By filtering approximately 92% of the literature, the analysis focuses on high-quality studies with a greater scientific impact.

The application of the classification rules and the analysis of the repetitiveness of the *source words* ensure that the SOTA is based on relevant and up-to-date research, enhancing the reliability and depth of the theoretical analysis. However, the classification rules can be adjusted to meet the specific needs of each research project, as well as the number of *source words* and/or knowledge areas found in each article to determine their relevance.

The research was deemed satisfactory, particularly regarding the accuracy of the methodology. This was demonstrated by the high F_1 Score achieved in discarding non-relevant articles, which outperformed similar approaches that utilize large language models (LLMs), such as the work reported by Alshami et al. (2023). While their approach achieved an F_1 Score of 0.80 in identifying "not related" classes, our methodology reached an F_1 Score of 0.83 for the task of discarding non-relevant articles. Although these results are promising, further work is required to conduct more accurate comparisons, as the studies differ in discipline. Therefore, future work should focus on cross-domain validation to enhance the robustness of the findings. Despite these positive outcomes, several areas for improvement were identified, underscoring the limitations of the framework. These include the following:

1. The framework's performance depends on the quality of the input data, which is influenced by the completeness and structure of the metadata or abstracts in the repositories. Inaccurate or poorly structured data may negatively impact system performance.
2. The accuracy of the system show that irrelevant documents could occasionally be retrieved, leading to unrelated information. To address this, an additional manual filtering phase, conducted by a panel of experts, is necessary.
3. Manual extraction of information is still required for subscription-based articles, even though APIs can automate the process, as seen with the arXiv repository. However, subscription-based repositories like IEEE Xplore, WoS, and Scopus require institutional access.

Future work will also focus on including additional repositories, such as bioRxiv, chemRxiv, medRxiv, PsyArXiv, and HAL, to enhance generalization across different areas. It will also aim to improve semantic extraction by implementing natural language processing (NLP) using open-source multimodal large language models (LLMs) and compare performance with related approaches.

5. Conclusions

Effective knowledge management is crucial in today's data-driven world, especially in the acquisition and transmission of information. However, the acquisition process is often slow and error-prone, limiting the development of state-of-the-art (SOTA) knowledge. This paper introduces a framework that automates the keyword search in article abstracts, using an algorithm that identifies the most relevant articles based on the frequency of key terms, referred to as *source words*. Unlike previous studies that explored automation in the systematic review of scientific literature using artificial intelligence, our methodology distinguishes itself by incorporating more efficient indexing through the generation of key terms and the systematic organization of documents.

A case study identified several knowledge areas with relevant *source words* for literature selection. The database was sourced from both subscription-based scientific repositories and the open-access arXiv repository via API. Six classification rules based on impact factor (IF), open access (OA), and JCR journal rankings were applied. The knowledge areas of colorimetry, food and beverages, and sensory analysis were the most represented. Some rules resulted in fewer articles, suggesting that certain classification criteria are more influential than others. The results highlight the multidisciplinary nature of the selected articles. This study ensures that the SOTA is based on relevant and up-to-date research by applying classification rules and analyzing *source words*. The rules can be customized to suit different research needs, adjusting the number of *source words* and knowledge areas for relevance.

References

Abdin, A. Y., & De Pretis, F. (2024). Measuring open access uptake: Methods and metrics to assess a market transformation. *Publications*, *12*(4), 33. [CrossRef]

Ali, A., Tufail, A., De Silva, L. C., & Abas, P. E. (2024). Innovating patent retrieval: A comprehensive review of techniques, trends, and challenges in prior art searches. *Applied System Innovation*, *7*(5), 91. [CrossRef]

Alkhamissi, B., Verma, S., Yu, P., Jin, Z., Asli, C., & Diab, M. (2023). OPT-R: Exploring the role of explanations in finetuning and prompting for reasoning skills of large language models. In *Proceedings of the 1st workshop on natural language reasoning and structured explanations (NLRSE)* (pp. 128–138). Association for Computational Linguistics. [CrossRef]

Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E. E., & Zayed, T. (2023). Harnessing the power of ChatGPT for automating systematic review process: Methodology, case study, limitations, and future directions. *Systems*, *11*(7), 351. [CrossRef]

arXiv API Access. (2025). *arXiv API access*. Available online: https://info.arxiv.org/help/api/index.html (accessed on 2 February 2025).

Asociación Española de Normalización y Certificación. (1991). *Information and documentation—Criteria for the assessment of information retrieval systems* (UNE 50-121-91). Asociación Española de Normalización y Certificación.

Cleverdon, C. W. (1984). Optimizing convenient on-line access to bibliographic databases. *Information Services & Use*, *4*, 37–47.

de la Torre-López, J., Ramírez, A., & Romero, J. R. (2023). Artificial intelligence to automate the systematic review of scientific literature. *Computing*, *105*, 2171–2194. [CrossRef]

Golinelli, D., Nuzzolese, A. G., Sanmarchi, F., Bulla, L., Mongiovì, M., Gangemi, A., & Rucci, P. (2022). Semi-automatic systematic literature reviews and information extraction of COVID-19 scientific evidence: Description and preliminary results of the COKE project. *Information*, *13*(3), 117. [CrossRef]

Hachicha, W., & Ghorbel, A. (2012). A survey of control-chart pattern-recognition literature (1991–2010) based on a new conceptual classification scheme. *Computers & Industrial Engineering*, *63*, 204–222. [CrossRef]

Ikarashi, A. (2023). Japanese research is no longer world class—Here's why. *Nature*, *623*, 14–16. [CrossRef]

International Organization for Standardization. (1985). *Information and documentation—Methods for assessing the quality of information retrieval systems* (ISO 5963:1985). International Organization for Standardization.

Jaiswal, A., Kumar, M., Pathak, A. R., & Yigzaw, K. Y. (2024). FaN-REMs: Fair and normalized retrieval evaluation metrics for learning retrieval systems. *IEEE Access*, *12*, 195370–195395. [CrossRef]

Jiang, S., Evans-Yamamoto, D., Bersenev, D., Palaniappan, S. K., & Yachie-Kinoshita, A. (2024). ProtoCode: Leveraging large language models (LLMs) for automated generation of machine-readable PCR protocols from scientific publications. *SLAS Technology*, *29*, 100134. [CrossRef]

Joseph, M. H., & Ravana, S. D. (2024). Reliable information retrieval systems performance evaluation: A review. *IEEE Access*, *12*, 51740–51751. [CrossRef]

Kirby, A. (2023). Exploratory bibliometrics: Using VOSviewer as a preliminary research tool. *Publications*, *11*(1), 10. [CrossRef]

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 195. [CrossRef]

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Miao, F., & Holmes, W. (2023). *Guidance for generative ai in education and research*. UNESCO. [CrossRef]

Morales-Alarcón, C. H., Bodero-Poveda, E., Villa-Yánez, H. M., & Buñay-Guisñan, P. A. (2024). Blockchain and its application in the peer review of scientific works: A systematic review. *Publications*, *12*(4), 40. [CrossRef]

Musslick, S., Bartlett, L. K., Chandramouli, S. H., Dubovad, M., Gobet, F., Griffiths, T. L., & Holmes, W. R. (2025). Automating the practice of science—Opportunities, challenges, and implications. *Proceedings of the National Academy of Sciences USA*, *122*(5), e2401238121. [CrossRef]

Ofori-Boateng, R., Aceves-Martins, M., Wiratunga, N., & Moreno-Garcia, C. F. (2024). Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: A comprehensive review. *Artificial Intelligence Review*, *57*, 200. [CrossRef]

Panayi, A., Ward, K., Benhadji-Schaff, A., Ibanez-Lopez, A. S., Xia, A., & Barzilay, R. (2023). Evaluation of a prototype machine learning tool to semi-automate data extraction for systematic literature reviews. *Systematic Reviews*, *12*, 187. [CrossRef]

Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., & Manning, C. D. (2024, May 7–11). *RAPTOR: Recursive abstractive processing for tree-organized retrieval*. The Twelfth International Conference on Learning Representations, Vienna, Austria.

Schick, T., & Schütze, H. (2021). It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 2339–2352). Association for Computational Linguistics.

Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S., & Wang, Y. (2024). An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, *12*, e55318. [CrossRef]

Turgel, I. D., & Chernova, O. A. (2024). Open science alternatives to scopus and the web of science: A case study in regional resilience. *Publications*, *12*(4), 43. [CrossRef]

Vatsal, S., & Dubey, H. (2024). A survey of prompt engineering methods in large language models for different NLP tasks. *arXiv*. [CrossRef]

Wu, S., Zhao, X., Yu, T., Zhang, R., Shen, C., Liu, H., Li, F., Zhu, H., Luo, J., Xu, L., & Zhang, X. (2021). Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning. *arXiv*. [CrossRef]

Zamiri, M., & Esmaeili, A. (2024). Methods and technologies for supporting knowledge sharing within learning communities: A systematic literature review. *Administrative Sciences*, *14*(1), 17. [CrossRef]