Software-Based Collection and Classification of Scientific Papers: A Use Case in Quantum Optics Research

Serhii D. Prykhodchenko, Oksana Yu. Prykhodchenko, Andrii A. Kolb, Dmytro V. Babets, Marcin Paszkuta, Krzysztof A. Cyran

Abstract - This paper presents research on software-based tools for the semi-automated collection and classification of scientific papers focusing on Quantum Optics research. The tool integrates keyword-based search, Portable Document Format extraction, and parallel processing using "term frequency-inverse document frequency" and the all-MiniLM-L6-v2 model for semantic analysis. It generates numerical similarity estimates, enabling efficient navigation and prioritization of sources. The modular design allows flexible integration of similarity algorithms. Despite challenges with dynamic anti-scraping mechanisms, the tool demonstrates significant potential in streamlining literature reviews. Future improvements include advanced NLP techniques and addressing ethical considerations to enhance accuracy and compliance.

Keywords - sources review, quantum optics, semantic similarity analysis, natural language processing (NLP).

I. INTRODUCTION

Conducting a thorough literature review is essential in computer science research to establish a solid foundation of knowledge. It helps identify the current state of the problem, explore existing approaches and technologies, and avoid duplicating prior work. This is particularly important in the field of optical quantum research, which generates a significant number of research papers. Analysing relevant literature enables researchers to define clear objectives, formulate hypotheses, assess task complexity, and select the most effective methodologies. Additionally, it provides insights into potential limitations and challenges, minimising risks during implementation. Ultimately, a well-prepared review accelerates the research process and enhances the significance of scientific outcomes.

ResearchGate is a unique platform for scientists and researchers. It allows users to publish articles, theses, and other scientific works while tracking their citations. The platform facilitates discussions, feedback, and knowledge exchange within the community. It also helps users discover cutting-edge research, stay updated on trends, and access the latest developments in their field, making it a powerful tool for advancing science and professional growth.

ResearchGate offers significant advantages for literature reviews in computer science. First, it aggregates a vast collection of scientific publications, enabling quick access to relevant articles and studies. Advanced search and filtering tools help efficiently organise materials by keywords, authors, and dates. Second, citation statistics and peer feedback provide insights into the impact and quality of sources. The platform often grants access to full-text articles, bypassing journal subscription barriers. Moreover, it fosters collaboration with other researchers, potentially yielding additional data or novel ideas. By leveraging ResearchGate, researchers can streamline the preparation process, deepen their analysis, and build a robust foundation for successful research.

The main idea of this study is to develop a tool for automated or semi-automated search and downloading of scientific articles and analysis of similarity of their textual content as applied to Quantum Optics Research. The key difference from existing software solutions is the generation of numerical estimates of the similarity of scientific papers, which allows researchers to navigate through sources and prioritise them quickly. Another unique feature is its modular design, which allows for easy integration or replacement of similarity algorithms. This flexibility increases the application's adaptability, allowing for deeper analyses or shifting the focus of similarity assessment depending on specific research needs.

II. STATE REVIEW

The study [1] presents LiteRev, an innovative tool designed to automate and accelerate the literature review process using advances in natural language processing (NLP) and machine learning (ML). Researchers often face challenges when dealing with large volumes of academic content; LiteRev solves this problem by quickly summarising and extracting relevant information from large datasets. The study highlights a descriptive evaluation of LiteRev's performance, demonstrating its ability to identify key data points while maintaining accuracy and reducing manual work. The tool can be adapted for various academic and professional settings, highlighting its versatility and potential for widespread adoption. Through systematic evaluation, the study highlights the effectiveness and efficiency of LiteRev in helping researchers make informed decisions faster. This significantly contributes to the growing integration of artificial intelligence into academic research, demonstrating a transformative approach to research resource management.

S.D.Prykhodchenko is with the Dnipro University of Technology, Dnipro, Ukraine (+380684054050; e-mail: prykhodchenko.s.d@nmu.one).

O.Yu.Prykhodchenko is with the Dnipro University of Technology, Dnipro, Ukraine (e-mail: prykhodchenko.o.yu@nmu.one)

A.A..Kolb is with the Dnipro University of Technology, Dnipro, Ukraine (e-mail: kolb.a.a@nmu.one)

D.V.Babets is with the Dnipro University of Technology, Dnipro, Ukraine (e-mail: babets.d.v@nmu.one)

Marcin Paszkuta is with the merQlab, Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology, Gliwice, Poland (e-mail: Marcin.Paszkuta@polsl.pl)

K. A. Cyran is with the merQlab, Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology, Gliwice, Poland (e-mail: krzysztof.cyran@polsl.pl)

The review paper [2] is devoted to scientific literature's current challenges and advances in automatic information extraction (IE). The authors emphasise that although text processing technologies have advanced significantly, their practical application is limited due to technical and logistical difficulties such as heterogeneous data formats and information volume. Considerable attention is given to opportunities for improving IE methods, especially for materials science, where automation can accelerate research. The conclusion presents directions for bridging the gap between theoretical developments and their implementation in scientific practice.

The paper [3] addresses challenges and solutions for automated data extraction from scientific web repositories. The authors note the lack of transparency in the metrics and algorithm repositories used to analyse publication activity, which limits researchers' capabilities. The paper presents an open-source tool that enables programmatic data extraction and scientific metrics, emphasising ethical considerations and the need to obtain consent from repository operators. The article [4] focuses on methods for automated collection and analysis of scientific literature from web repositories such as Google Scholar. The authors describe approaches for transforming unstructured data into structured formats, ensuring compatibility with secure protocols (HTTPS) and dynamic technologies (Ajax). A locality-sensitive hashing algorithm selects the most relevant publications, improving content sampling accuracy. The paper emphasises the importance of ethical and legal aspects, providing researchers and librarians with practical tools for efficiently retrieving scientific information.

The paper [5] analyses current methods for automating and semi-automating data extraction for systematic reviews. The authors review various approaches, including machine learning, natural language processing (NLP), and parsers, to speed up and simplify the data collection and analysis. Particular attention is given to tools that allow the integration of automated techniques into existing workflows while maintaining high accuracy and reliability of retrieval. The study highlights that despite significant progress, full automation remains a challenge, and further developments should focus on improving the interaction between researchers and automated systems.

A key study in this field is the systematic literature review by Salazar-Reyna et al., which highlights the increasing significance of data science and machine learning across various domains, particularly in healthcare engineering systems [6]. This review demonstrates how systematic methods can be utilized to collect and analyse relevant literature, offering a framework for understanding the role of data analytics in scientific research. Likewise, Tegegne et al. outline the methodologies applied in software development, stressing the importance of organized data collection from multiple databases like IEEE and Scopus, which are vital for accessing pertinent studies in software engineering [7]. Such a structured approach is critical for achieving thorough literature coverage.

Rivest et al. evaluate various methods for article-level classification, such as deep learning and bibliographic coupling, emphasizing the difficulties caused by the lack of gold-standard datasets in bibliometrics [8]. Their research underscores the importance of reliable evaluation metrics, like the Herfindahl index, to measure the performance of classification systems. This perspective is reinforced by Daradkeh et al., who introduce a convolutional neural network (CNN) framework for scientometric analysis, tackling the shortcomings of conventional classification methods that frequently result in low accuracy [9]. Their study demonstrates how deep learning techniques can improve the accuracy of classification tasks, particularly in the face of the rapidly growing volume of scientific literature.

Additionally, the study by Torres et al. on automatic result identification in software engineering papers sheds light on the challenges posed by unstructured data, emphasizing the need to develop novel methodologies to enhance classification accuracy [10]. This is consistent with the findings of Eykens et al., who examine fine-grained classification of social science journal articles through supervised machine learning techniques, illustrating the flexibility and applicability of these methods across a wide range of disciplines [11]. The use of machine learning techniques, including support vector machines (SVM), for text processing and classification of scientific papers has been investigated by Al-Habib et al. [12]. They highlight the significance of categorizing research based on content to identify areas of excellence and guide development strategies. This methodology is supported by Iqbal et al. [13], who explore in-text citation analysis using natural language processing (NLP) and machine learning, showcasing the wide-ranging applications of automated classification techniques in scientific research. Together, these studies underscore the potential of machine learning in enhancing the organization and analysis of academic literature.

In summary, the literature demonstrates a strong trend toward adopting software-based approaches for gathering and categorizing scientific papers. The integration of findings from multiple studies highlights the central role of machine learning techniques, especially deep learning and support vector machines (SVM), in improving the accuracy and efficiency of classification tasks. As the scientific field continues to expand and diversify, sustained research and innovation in these methodologies will be essential to manage the increasing volume and complexity of academic literature effectively. This ongoing development is critical to ensuring researchers can efficiently navigate and utilize the vast and ever-growing scientific knowledge.

Platforms such as Google Scholar and ResearchGate are actively developing their algorithms to protect data, including against parsing and mechanisms to detect and block automated requests, making the automated search and download process a complex and dynamic task. These developments aim to prevent unauthorised data collection and enforce intellectual property rights. In such an environment, automated data collection software solutions, subject to all ethical and legal constraints, must constantly adapt to new challenges, bypassing CAPTCHAs, handling dynamically generated content, or mimicking human behaviour. This requires regular algorithm updates, integration of new technologies (e.g., machine learning to analyse changes in the structure of web pages), and

consideration of legal and ethical aspects. Thus, research on the development and improvement of automated data collection and analysis methods remains highly relevant, as it maintains the effectiveness of existing tools and promotes the development of new approaches that meet current requirements and challenges.

III. Materials and Methods

Due to the restrictions imposed by Google and ResearchGate on automatic searching and parsing, the software module for automatic collection and analysis of scientific information has to change frequently and adapt to the updated restrictions. Nevertheless, the common elements of this system remain the same, so let us describe them.

The software module for the automatic collection and analysis of scientific information is a complex system consisting of several interrelated components, each performing its unique function in information processing. The first component, the input subsystem, is designed to interact with the user. It allows the user to enter queries, define search parameters, and set criteria for retrieving data. This subsystem plays a key role at the initial stage, providing the basis for subsequent operations.

Next is the search subsystem, which is responsible for generating and sending queries to various search platforms. It parses the results, extracting relevant information from the search engine responses. This stage is critical for correct data filtering and pre-selection of materials that meet the user's requirements.

Once the parsing results are received, the download subsystem is activated to process the retrieved information. To avoid legal and ethical conflicts, this system now operates semi-automatically, providing the user-researcher with the search results but allowing him/her to select and download the retrieved literature independently.

The final stage is the analysis subsystem. It processes the downloaded materials, extracting useful information from them. This includes compiling dictionaries of frequently used words, calculating other statistical indicators of the text, and data clustering. The latter is performed using various algorithms that allow the grouping of materials based on their content similarity. This module structure makes it a powerful tool for automating tasks related to extracting, analysing and systematising scientific information.

Thus, the software module for automatic collection and classification of scientific texts consists of the following components (fig.1):

1. Input subsystem. Provides a dialogue with the user and the primary input of the information sought.

2. Search subsystem. Provides querying of search sites and parsing of results.

3. Download subsystem. Processes the results of parsing and downloads scientific papers found in the search.

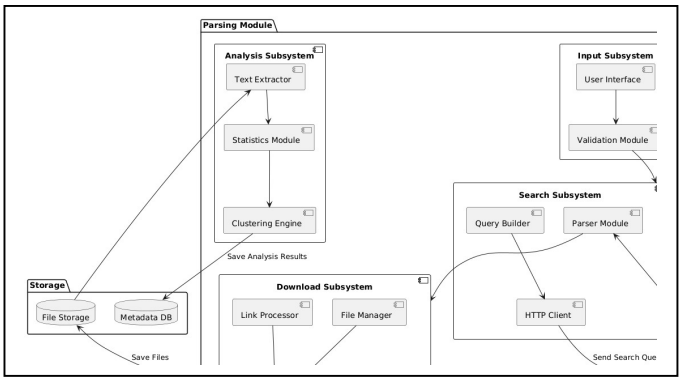

Figure 1. Software architecture model for the automatic collection and analysis of scientific information

In addition, service mechanisms such as storage subsystem and API for search queries can be used.

The pipeline (fig.2) for a software-based scientific sources collector and analyser starts by generating a keyword query to search for relevant PDF documents in Google Scholar. Once the target files are identified, their links are passed to the user for downloading. After downloading, the process is divided into parallel streams: the first stream uses the TF-IDF [14] method to extract key terms and analyse textual relevance, which allows the most relevant text fragments to be selected. The second stream uses the all-MiniLM-L6-v2 [15] model for semantic analysis and text vector representation, which provides a deeper understanding of the context and content of the documents. This approach allows combining traditional text processing methods with modern NLP technologies, ensuring high accuracy and efficiency of scientific literature analysis.

IV. Results

This section aims to demonstrate the performance of the proposed software tool in relation to the subject area of quantum optics. The main task is to identify thematic groups of scientific publications and evaluate their semantic proximity using the TF-IDF and all-MiniLM-L6-v2 models. This project's source code and associated materials are publicly accessible on GitHub at https://github.com/prykhodchenkosd/pdfreviewer. This repository provides all necessary resources, including code and sources, to support transparency, reproducibility, and collaboration. By sharing these materials, the project encourages further exploration and innovation within the community.

The first branch uses the tfidf function, which performs text document analysis using the TF-IDF method [14] and KMeans clustering [16]. It first loads text files, converts them into numerical vectors using TF-IDF (Term Frequency-Inverse Document Frequency) and extracts keywords.

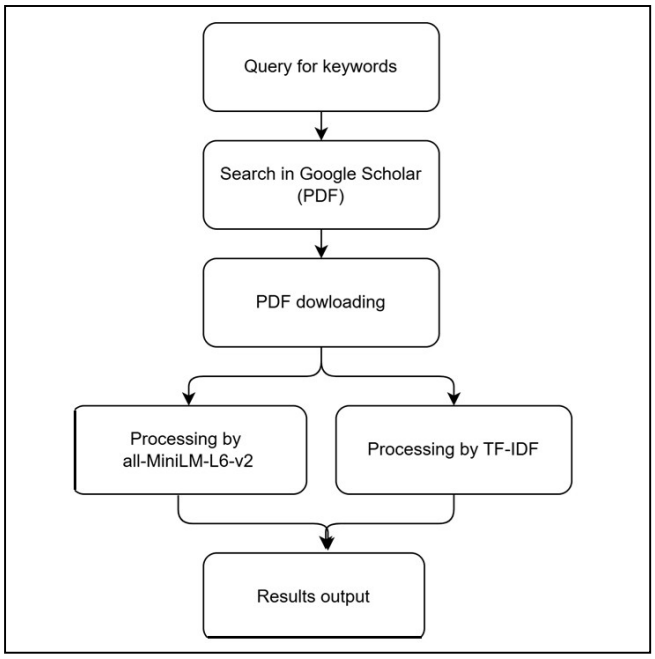

Figure 2. The pipeline for a software-based scientific sources collector and analyser

The TF-IDF method estimates the importance of words in a document relative to a collection of documents (corpus). It calculates the weight of a term by multiplying its frequency of occurrence in a document (TF) by the inverse of its frequency of occurrence in the whole corpus (IDF), which helps to identify words that are significant for a particular document but rarely occur in other documents in the collection.

$$TFIDF = \frac{n_t}{\sum_k n_k} \cdot log \frac{|D|}{|\{d_i \in D \vee t \in d_i\}|} \quad (1)$$

where n_t - is the number of occurrences of the word t in the document; $\sum_k n_k$ - is the total number of words in the document; $|D|$ - number of documents in the collection; $|\{d_i \in D \mid t \in d_i\}|$ - number of documents from collection D in which t occurs (when $n_t \neq 0$)

The program processes a collection of PDF documents by extracting and analysing their text content. It starts by iterating the PDF files in a specified directory, parsing each file to extract raw text using PDFParser and TextConverter. The text is then cleaned by removing stop words and short tokens, and the cleaned content is stored in separate text files. In addition, the feature generates word clouds to visualise frequently occurring terms and produces a list of the top 30 keywords for each document. These keywords are stored in a DataFrame for further analysis. The module of word cloud creation was used to obtain statistical indicators, which also provides an opportunity to visualise the content of the scientific work under consideration. The feature supports full processing and debug mode, allowing efficient testing and scaling.

KMeans clustering is then applied to group documents based on similarity. PCA is used for visualisation, which reduces the dimensionality of the data to two components, allowing a graph of the clusters to be plotted. The function saves the graph and outputs keywords and sample documents for each cluster, which is useful for analysing the thematic structure of the texts. Using the keywords 'quantum entanglement augmented reality', we downloaded 25 texts for analysis and obtained the clustering pattern shown in Fig. 3. The function hf_similarity_all, used in the second branch, computes the semantic similarity between text documents using the all-MiniLM-L6-v2 model from the Sentence Transformers library. The function encodes the texts into vector representations for each pair of documents. Then, the cosine similarity between each pair is computed, allowing us to estimate their semantic proximity degree.

The results of calculations also form a two-dimensional matrix of pairwise cosine similarities, which is used in further calculations. In addition, the results of this matrix can also form the basis for further work on materials related to the research field of optical quantum physics, namely, the values of cosine similarities can be used to obtain information about the most similar, from the point of view of the current algorithm, articles processed by the programme. An example of a part of such a table for the first 7 articles is shown in Table 1, where the maximum similarity is described as 1 and the minimum similarity as 0. As a result of processing this part, we can see that there is a similarity between articles 6 and 7, and articles 3 and 1 are maximally different in this example. For further visualisation it is convenient to transform pairwise cosine similarities by formula 2 to reverse meanings of similarities where "0" would mean "fully similar", and "1" would be "totally different".

$$CSim_{i,j} = 1 - CSim_{i,j} \quad (2)$$

where $CSim_{i,j}$ – cosine similarity value between article with number i and article with number j.

Fig. 3 shows a two-dimensional projection graph of clusters constructed using TFIDF. Colours indicate the belonging of documents to different thematic clusters. Analysis of keywords within clusters allowed us to identify the following dominant areas:

Cluster 1 (cyan): quantum mechanics and probability theory
Cluster 2 (red): quantum computing, cryptography and key distribution
Cluster 3 (light-green): quantum high-tech like IoT, AI, telecommunications, etc.
Cluster 4 (purple): fundamental research on quantum interference

The division into clusters turned out to be relevant and allows the user to quickly navigate the topics of the corpus. The results are written to a file, where their identifiers and similarity value are specified for each pair of documents. The function is useful for searching of similar scientific articles or grouping documents by subject. Next, the draw_sim function visualises the similarity matrix between documents using multidimensional scaling (fig.4).

Multidimensional scaling (MDS) is a dimensionality reduction technique that converts a matrix of pairwise distances or similarities between objects into a low-dimensional space (e.g. 2D or 3D) while preserving their relative distances. This allows complex multidimensional data to be visualised in a form that is easy to analyse, revealing hidden structures or clusters.

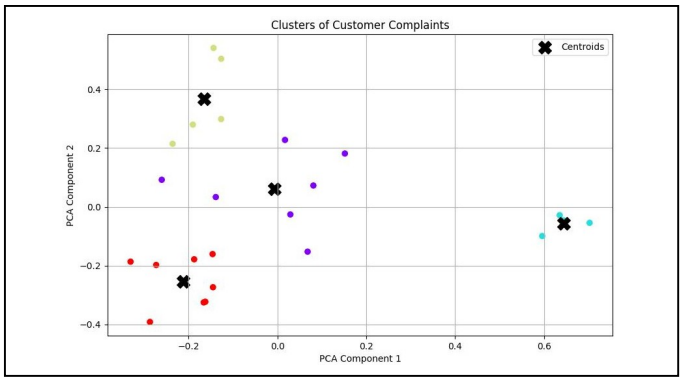

Figure 3. Clustering based on papers contents

MDS is then applied to reduce the dimensionality of the data to 2D and 3D, allowing relationships between documents to be visualised as points on the plane or in space. The points are annotated with document indices for easy interpretation. The feature is useful for analysing the structure of data, for example, when studying a semantic proximity of texts or document clustering.

Table 1 shows the cosine similarity coefficients between pairs of articles. For example, articles 6 and 7 showed a good degree of similarity (0.64), which corresponds to their thematic focus on quantum philosophy and cosmology. This indicates that the model works correctly and is able to find publications with similar content without taking into account superficial terms.

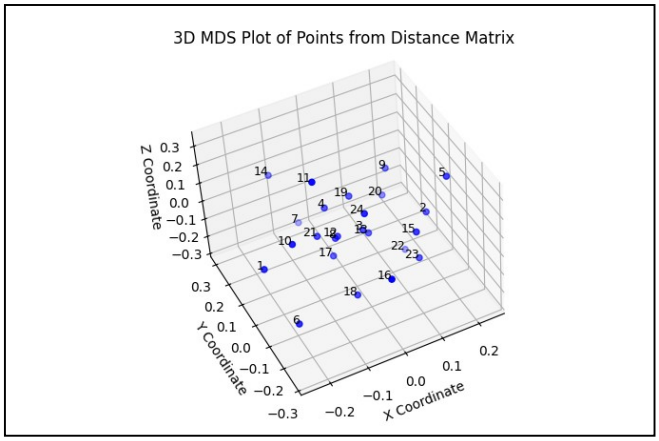

Figure 4. 3D-representation of the proximity of papers by content

TABLE I. PAIRWISE COSINE SIMILARITIES

Article number	Article number						
	1	*2*	*3*	*4*	*5*	*6*	*7*
1	1.0	0.23	0.10	0.28	0.39	0.33	0.44
2	0.23	1.0	0.20	0.18	0.39	0.39	0.40
3	0.10	0.20	1.0	0.35	0.28	0.14	0.13
4	0.28	0.18	0.35	1.0	0.29	0.19	0.23
5	0.39	0.39	0.28	0.29	1.0	0.46	0.40
6	0.33	0.39	0.14	0.19	0.46	1.0	0.64
7	0.44	0.40	0.13	0.23	0.40	0.64	1.0

V. ANALYSIS AND DISCUSSION OF RESULTS

The system was evaluated on a curated dataset of 25 scientific articles in the domain of quantum optics, extracted using keyword-based queries. The documents were preprocessed and embedded using TF-IDF and the all-MiniLM-L6-v2 model to capture both lexical frequency and semantic relationships. A dimensionality reduction technique (t-SNE) was employed to visualize the similarity space, and clustering was performed using k-means. The resulting visualizations revealed distinguishable groupings of documents, suggesting that the system is capable of identifying latent topical structures within the corpus. These results were corroborated through qualitative inspection of representative samples from each cluster.

The clusters that emerged were not only internally coherent but also mapped intuitively to established subfields in quantum optics, such as quantum key distribution, photonic sources, and quantum circuit implementations. This supports the idea that lightweight sentence transformers like MiniLM, when combined with lexical models like TF-IDF, are capable of uncovering semantically meaningful relationships even in highly technical domains. Notably, some article pairs—such as those shown in Table 1—exhibited strong cosine similarity, corresponding to near-identical research themes. This alignment between automated similarity measures and expert judgment indicates that the tool can be of practical value in supporting literature reviews and thematic mapping.

However, the system also displayed limitations that merit further attention. While the clusters were visually distinct, the interpretability of their content relied on manual keyword extraction and domain expertise. This points to a broader challenge in unsupervised semantic clustering: algorithmic success in segmentation does not automatically imply clarity in cluster labeling. Furthermore, although MiniLM offers speed and generalization, its embedding space is trained on general-domain corpora, which may underrepresent domain-specific jargon and nuances. This was evident in cases where articles with superficially different vocabulary but similar conceptual content were placed into separate clusters.

Another critical aspect relates to the broader applicability of the system in real-world academic workflows. While the current implementation functions effectively on pre-acquired PDF datasets, it faces scalability and integration challenges due to anti-scraping protections imposed by major academic repositories. The system's impact could be significantly enhanced by incorporating API-based retrieval from open-access databases (e.g., arXiv, Semantic Scholar) and by offering post-processing tools for interactive filtering and tagging. Moreover, future development should consider the integration of domain-specific pretrained language models such as SciBERT or SPECTER, which are better suited to capturing scientific discourse and could improve classification accuracy in specialized fields.

Overall, the results presented here demonstrate that even

with lightweight models and relatively simple techniques, it is possible to achieve meaningful semantic organization of scientific literature. The tool offers a promising foundation for more advanced systems that assist researchers in navigating large volumes of academic text. With improvements in model specialization and system integration, such tools could evolve into indispensable components of the modern scientific workflow.

VI. Conclusion

This work presents an integrated pipeline for automated scientific literature analysis, utilizing TF-IDF, semantic similarity models (e.g., all-MiniLM-L6-v2), and multidimensional scaling (MDS) for visualization. It addresses the challenge of processing large volumes of academic texts by enabling efficient identification, comparison, and clustering of relevant works. The combination of traditional text mining and modern NLP techniques results in a flexible system suitable for literature review and trend analysis.

The modular architecture supports the integration of alternative algorithms, while MDS-based visualizations offer intuitive insights into document relationships. However, the system's performance hinges on adaptability to evolving data access restrictions, such as anti-scraping measures on scholarly platforms.

Future improvements should target scalability, the refinement of similarity models through domain-specific fine-tuning, and consideration of ethical and legal data use. This research contributes a practical tool for streamlining academic exploration without compromising analytical rigor.

References

[1] Orel, Erol & Ciglenecki, Iza & Thiabaud, Amaury & Temerev, Alexander & Calmy, Alexandra & Keiser, Olivia & Merzouki, Aziza. (2023). An Automated Literature Review Tool (LiteRev) for Streamlining and Accelerating Research Using Natural Language Processing and Machine Learning: Descriptive Performance Evaluation Study. Journal of Medical Internet Research. 25. e39736. 10.2196/39736.

[2] 'Hong, Zhi & Ward, Logan & Chard, Kyle & Blaiszik, Ben & Foster, Ian. (2021). Challenges and Advances in Information Extraction from Scientific Literature: a Review. JOM. 73. 1-18. 10.1007/s11837-021-04902-9.

[3] Meschenmoser, Philipp & Meuschke, Norman & Hotz, Manuel & Gipp, Bela. (2016). Scraping Scientific Web Repositories: Challenges and Solutions for Automated Content Extraction. D-Lib Magazine. 22. 10.1045/september2016-meschenmoser.

[4] Ahmed A., Khan M.A., Ishtiaq A. 'Web Scraping for Scientific Discovery: Strategies for Secure Data Retrieval, Structured Transformation, and Relevant Content Selection." IEEE-SEM, Volume 11, Issue 10, October 2023 ISSN 2320-9151. https://www.ieeesem.com/researchpaper/Web_Scraping_for_Scientific_Discovery_Strategies_for_Secure_Data_Retrieval_Structured_Transformation_and_Relevant_Content_Selection.pdf.

[5] Schmidt L, Finnerty Mutlu AN, Elmore R, Olorisade BK, Thomas J, Higgins JPT. Data extraction methods for systematic review (semi)automation: Update of a living systematic review. F1000Res. 2021 May 19;10:401. doi: 10.12688/f1000research.51117.2. PMID: 34408850; PMCID: PMC8361807.

[6] Salazar-Reyna, R., Aleu, F. G., Granda-Gutierrez, E. M., Díaz-Ramírez, J., Garza-Reyes, J. A., & Kumar, A. (2020). A systematic literature review of data science, data analytics and machine learning applied to healthcare engineering systems. Management Decision, 60(2), 300-319. https://doi.org/10.1108/md-01-2020-0035

[7] Tegegne, E. W., Seppänen, P., & Ahmad, M. O. (2019). Software development methodologies and practices in start-ups. IET Software, 13(6), 497-509. https://doi.org/10.1049/iet-sen.2018.5270

[8] Rivest, M., Vignola-Gagné, É., & Archambault, É. (2021). Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling. PLOS ONE, 16, e0251493. https://doi.org/10.1371/journal.pone.0251493

[9] Daradkeh, Mohammad & Abualigah, Laith & Atalla, Shadi & Mansoor, Wathiq. (2022). Scientometric Analysis and Classification of Research Using Convolutional Neural Networks: A Case Study in Data Science and Analytics. Electronics. 11. 2066. 10.3390/electronics11132066.

[10] Torres, J. A. S., Cruzes, D. S., & Salvador, L. d. (2012). Automatic results identification in software engineering papers. is it possible?. 2012 12th International Conference on Computational Science and Its Applications. https://doi.org/10.1109/iccsa.2012.27

[11] Eykens, J., Guns, R., & Engels, T. (2021). Fine-grained classification of social science journal articles using textual data: a comparison of supervised machine learning approaches. Quantitative Science Studies, 2(1), 89-110. https://doi.org/10.1162/qss_a_00106

[12] Al-Habib, H., Imah, E. M., Puspitasari, R. D. I., & Prahanı, B. K. (2023). Text processing using support vector machine for scientific research paper content classification. Advances in Intelligent Systems Research, 273-282. https://doi.org/10.2991/978-94-6463-174-6_20

[13] Iqbal, S., Hassan, S., Aljohani, N. R., Alelyani, S., Nawaz, R., & Bornmann, L. (2021). A decade of in-text citation analysis based on natural language processing and machine learning techniques: an overview of empirical studies. Scientometrics, 126(8), 6551-6599. https://doi.org/10.1007/s11192-021-04055-1

[14] Jones K. S. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation: journal. — MCB University: MCB University Press, 2004. — Vol. 60, no. 5. — P. 493—502. — ISSN 0022-0418.

[15] Chen Yin and Zixuan Zhang. "A Study of Sentence Similarity Based on the All-minilm-l6-v2 Model With 'Same Semantics, Different Structure' After Fine Tuning." Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024), Atlantis Press, 2024, pp. 677-684. DOI: 10.2991/978-94-6463-540-9_69.

[16] Lloyd, Stuart P. (1957). "Least square quantization in PCM". Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, Stuart P. (1982). "Least squares quantization in PCM" (PDF). IEEE Transactions on Information Theory. 28 (2): 129–137. CiteSeerX 10.1.1.131.1338.