London Review of Education

LONDON REVIEW OF EDUCATION

UCLPRESS

Special issue: ***Systematic reviews in education***

Research article

Uses of artificial intelligence and machine learning in systematic reviews of education research

Henrik Karlstrøm[1,*]

[1] Nordic Institute for Studies of Innovation, Research and Education (NIFU), Oslo, Norway

* Correspondence: henrik.karlstrom@nifu.no

Submission date: 6 October 2023; Acceptance date: 24 October 2024; Publication date: 4 December 2024

How to cite

Karlstrøm, H. (2024) 'Uses of artificial intelligence and machine learning in systematic reviews of education research'. *London Review of Education*, 22 (1), 40. DOI: https://doi.org/10.14324/LRE.22.1.40.

Peer review

This article has been peer-reviewed through the journal's standard double-anonymous peer-review process, where both the reviewers and authors are anonymised during review.

Abstract

The speed and volume of scientific publishing is accelerating, both in terms of number of authors and in terms of the number of publications by each author. At the same time, the demand for knowledge synthesis and dissemination is increasing in times of upheaval in the education sector. For systematic reviewers in the field of education, this poses a challenge in the balance between not excluding too many possibly relevant studies and handling increasingly large corpora that result from document retrieval. Efforts to manually summarise and synthesise knowledge within or across domains are increasingly running into constraints on resources or scope, but questions about the coverage and quality of automated review procedures remain. This article makes the case for integrating computational text analysis into current review practices in education research. It presents a framework for incorporating computational techniques for automated content analysis at various stages in the traditional workflow of systematic reviews, in order to increase their scope or improve validity. At the same time, it warns against naively using models that can be complex to understand and to implement without devoting enough resources to implementation and validation steps.

Keywords literature reviews; machine learning; bibliometrics; computational text analysis

Introduction

The field of education research is increasingly affected by socio-technical challenges. In recent years, a combination of increased digitisation of educational offerings and world-spanning crises such as the global Covid-19 pandemic have had profound consequences both for the surrounding support structures for, and for the content of, educational research. Topics in education research have moved towards a concern with digitisation, psychological and medical factors and the impact of these on curriculum development and instruction (Polat, 2022). There is also rising interest in questions of inclusivity in education, particularly in situations of social change (Pak and Ravitch, 2021). This context, coupled with an increasing demand for rapid dissemination of empirical evidence in times of crisis (Gorbea Díaz et al., 2023), means that the conditions for systematic appraisals of new research in the field have shifted. Simultaneously, the landscape of scientific publishing has undergone dramatic changes in the past few decades, both in terms of the volume of publications and in new forms of dissemination and the emergence of new fields and subfields in most disciplines. These changes have a bearing on the practice of reviewing and summarising large corpora of academic texts.

Even as changing conditions bring new challenges, there are developments within the practice of systematic knowledge summarisation which might help meet these. Concurrently with the evolution of the publishing landscape, new developments in the capabilities of machine-assisted analyses of lexical and semantic content ('ML/AI techniques' from now on) have given rise to new methods for conducting large-scale review and summarisation. In fields with a high degree of standardisation in reporting results, such as medicine, the use of ML/AI techniques in research synthesis is already common (Marshall and Wallace, 2019; Van Dinter et al., 2021). The practice is also spreading to other quantitatively oriented fields where standardised protocols for statistical meta-analysis can be developed (Ioannidis, 2022). In fields with a higher degree of heterogeneity in reporting practices, such as education research, the use of ML/AI techniques for textual analysis is still limited, although growing (Ayanwale et al., 2024).

If used properly, ML/AI techniques can present one way to at least partially address new challenges arising from the intensification of academic publishing. At the same time, it is important to be aware of the trade-offs that come with increased automation of text analysis, particularly in terms of validity and trust in the results.

In this article, I will discuss how new computational techniques can assist in all phases of the systematic review process, from text retrieval and screening to analysis of the content within publications using machine learning and contextual analysis of the relations between documents using bibliometric methods. The article does not present original research using these tools. Instead, it aims to provide: (1) a description of the current challenges and opportunities presented by the rise of ML/AI techniques in systematic reviews related to the field of education; (2) a typology of review tasks where such techniques can be used; and (3) an appraisal of the trade-offs inherent in the adoption of these techniques. I present a set of promising avenues for the automation of manual tasks that are proving untenable when met with corpora above a certain size. This avoids having to limit the number of publications eligible for reviews without first considering their fit for the review. Rather than supplanting the expert assessments of reviewers, the aim is to provide reviewers with a solid conceptual foundation for understanding the parts of the review process that can be supplemented by quantitative methods, and which considerations must be taken when sampling, filtering, mapping and summarising research fields.

Challenges in systematic reviews

In this section I will briefly discuss two of the main challenges facing systematic publication analysis today: the explosive growth in scientific literature and the increased fragmentation of scientific fields. These have developed against a backdrop of changing expectations from policymakers and society at large for more comprehensive and more rapidly produced reviews of relevant research in the face of disruptive events such as the recent global Covid-19 pandemic (W.-T. Wang and Wu, 2021). Increased demand and shorter turnaround place an onus on the systematic reviewer to combine more efficient methods with rigorous quality control to ensure reliability in their work (Buhagiar and Anand, 2023).

Explosive growth in publications

In the latter half of the twentieth century and well into the twenty-first century, the global output of research articles has been doubling roughly every fifteen years (Bornmann et al., 2021; Thelwall and Sud, 2022). Such rapid expansion of the scientific corpus has serious implications for systematic reviews, especially in what might be called text-interpretative fields such as education research, where literature is highly heterogeneous in form and content, and dispersed across more numerous but smaller journals (Bearman et al., 2012). Traditional methods of literature analysis, often involving manual sifting and vetting of articles, become increasingly untenable as the volume of publications continues to rise.

The sheer volume of publications poses a logistical challenge to review projects that rely on manual identification and classification of publications for inclusion. The main problem, however, is the increased risk of low validity that results from attempting to implement stringent search constraints to limit eligible search results (Cian, 2021). Missing out on pivotal studies can compromise the integrity and findings of the review. As the volume of scientific outputs mushrooms, ensuring thorough validation becomes difficult (Lefebvre et al., 2019). A vast corpus means a greater number of studies to scrutinise, methodologies to understand and results to interpret. It is imperative, therefore, to consider how systematic reviews can maintain rigour, depth and breadth facing such an abundance of potentially relevant information.

Field fragmentation

The proliferation of scientific literature has been accompanied by the increasing fragmentation and specialisation of scientific fields (Sjögårde and Ahlgren, 2020; Q. Wang and Waltman, 2016). While indicative of the maturation and refinement of scientific disciplines, this presents substantial challenges for systematic reviews. Concurrent with this is an increased focus on interdisciplinary collaboration, resulting in more collaborative publishing across fields (Glänzel and Debackere, 2022). This trend is present in education research, which exhibits tendencies towards both increased fragmentation and multidisciplinarity (Huang et al., 2020).

Frequently, systematic reviews are undertaken to gain an understanding of questions that cut across disciplinary boundaries. Synthesising insights across multiple sub-disciplines requires generalised knowledge of the process of systematic reviews combined with domain-specific knowledge (Park et al., 2021). There is a constant tension between depth and breadth. Ensuring a comprehensive review across several fragmented subfields often means wading through disparate terminologies, methodologies and even epistemologies, which can be an arduous and intricate task.

Expecting reviewers to possess deep expertise across all relevant sub-disciplines covered by a systematic review is in many cases unrealistic. This limitation raises questions about the proficiency with which reviewers can validate findings from subfields outside their core expertise. Nuances in methodologies, terminologies or theoretical frameworks that are specific to a sub-discipline might be misconstrued or oversimplified by someone unfamiliar with that specialisation (Shahjahan et al., 2022). This is particularly true in cases where the literature under review has fewer standardised reporting elements and data suitable for rigorous meta-analysis (M. Campbell et al., 2020; Tong et al., 2012), such as is the case with education research.

Possibilities in publication analysis

The challenges posed by the growth and diversification of scientific literature underscore the need for expanding the toolbox of systematic review in literature retrieval (Gusenbauer and Haddaway, 2020), relevance filtering (Rethlefsen et al., 2021) and content summarisation (El-Kassas et al., 2021). Traditional search methods, reliant on keyword-based querying and manual filtering, are becoming less feasible and efficient, given the vastness and complexity of today's academic databases (Harari et al., 2020).

However, the rise and fragmentation of scientific publishing is not the only significant trend seen in the past decades. In the same period, there have been major advances in the development of computer-assisted tools for handling the content within, and the relations between, text documents (Khurana et al., 2023). These developments open new opportunities for the systematic analysis of scientific texts, aided by better language models and better access to computing hardware and literature metadata. Together, machine-learning techniques and relational bibliometric analysis can alleviate some of the pains of attempting systematic synthetisation of the research literature (Pan et al., 2024), and potentially reduce the effects of human error in various parts of the review process (Bannach-Brown et al., 2019; Kusa et al., 2023).

Natural language processing

ML/AI techniques present promising solutions to the challenges that have long plagued the systematic review process, offering both enhanced efficiency and depth of analysis. Whereas computational text analysis techniques were somewhat esoteric and highly specialised fields twenty years ago, three developments have combined to make ML/AI techniques available to researchers outside computer science or specialising in language-processing tasks: text models have vastly improved; the cost and task complexity of text analysis has dropped; and mature software support systems have appeared.

First, there have been clear improvements in text models. Unlike earlier models that relied heavily on manual feature engineering and could only capture surface-level patterns (Raffel et al., 2020), contemporary models such as transformers can understand context, semantics and even nuances in texts (Min et al., 2021). The ability of modern language models to consider contextual information makes them adept at tasks such as identifying sentiment in text (Wankhade et al., 2022), identifying and classifying topics (Vayansky and Kumar, 2020) and clustering documents based on their semantic content (Ghosal et al., 2020). For systematic reviews, this can translate into more accurate literature categorisation, richer extractions of insights and even the potential to identify overarching themes across disparate studies.

Second, the past decade has seen tremendous growth in specialised hardware and software designed to handle large-scale text-processing tasks (Lauriola et al., 2022). Assuming the technical competency is there, even vast corpora can be processed locally, reducing dependency on costly cloud services or high-end data centres.

Third, the support systems for doing ML and natural language processing (NLP) analysis have matured over the same period (Hewage and Meedeniya, 2022). The ML and NLP landscape is defined not only by its algorithms and hardware, but also by the ecosystems that support them. There has been a proliferation of user-friendly software tools tailored for text analysis (Gkevrou and Stamovlasis, 2022; Qi et al., 2020) and off-the-shelf solutions providing pre-trained models and easy-to-use application programming interfaces (APIs) (Gamieldien, 2023). Researchers with access to the right technical competencies can also train their own models with open-source access to the underlying language models (Wang et al., 2024). Extensive support documentation and training material is readily available. Together, these developments point towards a maturation point for the inclusion of ML and NLP techniques in their review workflow.

Stages of a systematic review

To better understand how computational techniques can fit into well-established workflows for systematic reviews, it helps to understand the distinct stages of the review process. The rest of this article will describe the review process, where computational techniques can be employed in such a workflow, and new challenges that may arise from the use of automation that reviewers must be aware of and able to answer satisfactorily.

We can divide the workflow of reviews into four distinct phases (Newman and Gough, 2020):

1. Operationalisation of research questions and conceptual framework
2. Identification of potentially relevant literature and document retrieval
3. Analysis and summarisation of the content of publications
4. Analysis and visualisation of metadata and content.

In this article, I focus on the last three stages, as the operationalisation and conceptualisation steps criteria lean heavily on domain expertise, and they are still reliant on manual design decisions.

Identifying relevant publications

Most systematic reviews start their literature identification and retrieval phase with a keyword search, using the resulting publication set either for bounding the corpus or as a starting point for various forms of snowballing and/or corpus supplement strategies (Polanin et al., 2017). Within the context of structured databases such as Web of Science or Scopus this will continue to be the most common method, meaning there is little scope for computational techniques to play a large role in this step in the process.

However, most keyword-based search techniques result in a large share of publications of low relevance to the review topic or research question. Some review tasks start with a corpus of publications connected through other criteria than topical or field similarity. This means that being able to quickly assess large numbers of publications for eligibility or clustering and classification can provide major benefits, especially when the corpus size expands beyond what is feasible to manually handle.

For example, rather than rely on relevancy criteria defined through the search strategy (for example, only publications from a certain geographic area, or from a very limited time period), computational techniques can be used to exclude or include publications based on criteria related to relational (De Bellis, 2009) or semantic characteristics (Van de Schoot et al., 2021) of the publications.

Expanding beyond the citation signal, NLP techniques can be employed to match publications based on lexical patterns and semantic content (Chandrasekaran and Mago, 2022). This ensures that even articles that do not explicitly use the predefined keywords but discuss the topic in question or adjacent, pertinent topics, have a chance of being captured. Making use of such techniques can also improve recall of document retrieval (Kuzi et al., 2020).

Other techniques involve text classification algorithms for relevancy scoring. Systematic reviews in education research have made use of such algorithms when they have been predefined and implemented in existing review software such as Leximancer (Thomas, 2014), Covidence (Jackson et al., 2022) and Rayyan (Bhatti et al., 2023), but the ability to fine-tune models for sensitivity towards domain-specific terms has been shown to yield good results in the field of education (Z. Liu et al., 2023). The most basic method is to use ML models trained to classify texts based on predefined relevancy criteria. By feeding these models a training set of relevant and non-relevant articles, they can learn to discern the characteristics of pertinent publications. Once trained, they can process large volumes of literature, efficiently categorising them as relevant or not. Automated classification speeds up the initial filtering process, reduces manual labour and ensures consistent application of relevancy criteria across a large corpus. However, validity concerns necessitate pre- and post-application manual validation of these techniques (Song et al., 2020), meaning that reductions in time and effort only manifest at larger scales.

More complex filtering techniques involve using a clustering or multiclass classification algorithm to identify clusters based on their semantic and topical similarities, to then identify sub-corpora of higher relevance for inclusion in the review process (Weisser et al., 2020). Similarly, experiments with large language models (LLMs) have shown good performance on clustering tasks (Keraghel et al., 2024).

Analysing and summarising the content of publications

After defining the set of eligible publications for a review and validating the resulting corpus, the next step in most review processes is analysing and summarising the content of the publications. Traditionally, this step could only be done by the reviewer reading and summarising the content in a manual fashion. The benefit of this is that human judgement can be attached to the resulting analysis, but the obvious drawback is that it scales very poorly with corpus size.

Computational text analysis offers far superior scalability. Modern NLP architectures using word embeddings or transformers have been shown to achieve human-level classification and summarisation scores, meaning human evaluators agree with the algorithmic classification about as often as they agree with other humans completing the task (Bird et al., 2023; Occhipinti et al., 2022). For some tasks, sentiment analysis can be used to understand the valence of a publication, particularly in identifying supporting or detracting citations to other literature (Wang et al., 2022).

In addition to identifying conceptual relationships through semantic similarity, some models can be used for automatic summarisation or data extraction tasks (Jethani et al., 2023; Wagner et al., 2022). LLMs, with their large context windows and fine-tuning for extractive tasks, offer a promising avenue for automated text summarisation (Bianchini et al., 2024; S. Liu et al., 2024). This is particularly useful when identifying specific sections of publications, for example, extracting descriptions of methodologies, or other clearly delimited summarisation tasks (de la Torre-López et al., 2023). For whole-document summarisation, current models have been shown to struggle with summarisation of long-form texts (El-Kassas et al., 2021), particularly if the task is of an abstractive (that is, generating new sentences that capture semantic meaning) rather than an extractive kind. This can be alleviated by introducing indicators of domain knowledge or additional metadata in the training process (Xie et al., 2022), but careful thought must go into integrating these techniques into the review workflow. Still, the largest providers of LLMs all currently provide ways to define sets of documents that can act as a knowledge base for the model, reducing their tendency for hallucinations and increasing validity of the summarisation (S. Liu et al., 2024).

The trade-off in using these techniques is ceding some control to the algorithms (Kasneci et al., 2023). It is usually possible to inspect the weighting scores of individual records in classification tasks, and some variable importance measures can be computed to identify which terms contribute the most to a particular classification. However, providing this in a meaningful way for thousands, if not tens of thousands, of publications can be challenging. The effect is that the reviewer will have to draw validity from the strength of the pre- and post-validation steps undertaken earlier in the process (Susnjak et al., 2024).

Metadata analysis and visualisation

Concurrent with content analysis, relational analysis can help in understanding the research context of a set of publications (F. Campbell et al., 2023). Situating research in time and place adds contextual information and can itself be used to identify clusters of researchers or topics. In many cases, one of the goals of the review process is to gain an understanding not only of what is covered in the corpus, but also of who is contributing.

The most common techniques of relational analysis are used for domain mapping, with the goal of mapping out the underlying structure of networked relations that can be inferred from the metadata. These include co-authorship networks, influence lineage through citation networks or mappings of publication channels for any given topic. Biographical metadata can be used to construct profiles of the academic milieux of the corpus, to understand geographical or institutional distribution (Higham et al., 2022; Rungta et al., 2022). Temporal network analysis can be used to trace the development of topics and scientific domains (Jiang and Liu, 2023; Vital and Amancio, 2022). In the overlap between relational and contextual analysis, topic modelling using title and abstract text has been shown to produce good results, albeit often requiring supervised training and manual validation to ensure good reconstruction of topics (Held et al., 2021).

While relational analysis is often used as a context-providing supplement to content analysis, its methods are more mature in terms of tested validity, and they offer more in the way of interpretability of results. Because of this, they can also serve as extra steps towards validation of the content analysis techniques. Using topic modelling in combination with text classification can be part of a comparative validation step. In addition, relational analysis lends itself well to visualisation. Network visualisations of citation, co-publication or topic similarity graphs offer a way to manually inspect and validate the output of the algorithms (Kossmeier et al., 2020). This has the potential to increase the validity of the review project.

The integration continuum

As should be evident from the discussion so far, there are multiple phases in the systematic review process where ML/AI techniques can be integrated. This integration is not a binary choice, but rather a continuum with varying degrees of implementation. One can envision moving along this spectrum, from minimal to full integration, depending on the complexity, size, goals and available resources of the project. As the use of computational methods intensifies in a project, scaling in terms of corpus size and analytical methods can be achieved, but not without incurring added costs in terms of project complexity and introducing extra validation steps. Integrating new techniques requires different skill sets and the ability to work in a cross-disciplinary fashion, both of which have project size and complexity costs related to them. Table 1 summarises the characteristics of the various degrees of integration and gives some hints as to when it makes sense to apply them.

Table 1. Characteristics and applicability of various computational integration modes

Integration	Characteristics	Applicability for
Minimal	Traditional search and manual selection remain dominant. Computational methods are applied in limited scenarios, such as initial eligibility filtering or to understand citation patterns of core articles.	• Smaller, narrowly focused reviews, or when the review team has limited familiarity with ML and bibliometric tools. • Topics that are highly specialised and require domain-specific interpretations.
Moderate	Traditional methods and computational techniques applied side-by-side, complementing each other. ML/NLP might be used for initial data cleaning and clustering, but in-depth analysis and summarisation remain manual. Bibliometrics might be used to identify key publications and authors, serving as a guide for manual exploration.	• Medium-sized projects, or when the literature spans multiple, interconnected domains.
Significant	Computational techniques play a guiding role in all phases, often iteratively, providing initial shortlists, flagging discrepancies or highlighting emerging themes. Critical decisions and final interpretations are made by human reviewers.	• Larger, more complex reviews, especially when the literature is vast and evolves rapidly.
Full	Almost every stage, from data gathering to analysis, is dominated by computational methods. ML/NLP models not only categorise and cluster, but might also provide preliminary interpretations or summaries.	• Very large-scale reviews, or when the objective is to provide a broad overview rather than a deep dive. • Projects with significant time constraints • Overview or simplified review projects where the goal is to get a rapid understanding of a vast body of literature.

The choice of where a review falls on this spectrum should be strategic and driven by the unique requirements and constraints of the project. While the allure of advanced computational techniques is undeniable, it is crucial to remember that the goal of a systematic review is to provide accurate, insightful and actionable analysis for policymakers. The tools employed, be they manual or computational, should always serve this primary objective.

Challenges with the integrated approach

One of the primary concerns with employing automated systems, particularly complex ML models, is the 'black box' nature of their operations (Yan et al., 2024). While ML models can efficiently process vast amounts of text and identify patterns beyond human capability, their decision-making processes can often be opaque (Tao et al., 2022). Additionally, while much work is done on testing models on various text analysis tasks, the field still lacks rigorous, transparent benchmarks for model evaluation (O'Connor et al., 2019). This lack of transparency poses challenges in the validation of the selection, classification and summarisation steps. If reviewers cannot understand or explain why certain texts were selected or categorised in a particular manner, it can lead to scepticism regarding the model's decisions. This opacity can thereby undermine the perceived validity and trustworthiness of the entire review process.

Systematic reviews traditionally rest on domain expertise, where review validity is based on the reviewer's expert assessment. As noted in the introduction, this can already pose a problem for more fragmented fields such as education research, where there is a higher heterogeneity in terminology used (Coe et al., 2021; Newman and Gough, 2020). The integration of quantitative text analysis introduces a technical dimension that might be alien to many reviewers. A review project must now introduce rigorous training, testing and validation cycles to the process, particularly for the more integrated procedures. The need to understand and sometimes tweak algorithms, validate model outputs or interpret complex network graphs can be daunting for those without a background in computational methods. This mismatch can result in a reluctance to adopt these tools or, worse, their misuse due to a lack of understanding.

Given the technical challenges of custom-building and maintaining ML/NLP models, using off-the-shelf software or proprietary platforms might be the only feasible road to integration. While these offer user-friendly interfaces and promise comprehensive analysis, they come with their own set of challenges. First, they can be costly, limiting access for researchers with constrained budgets. Second, proprietary systems further exacerbate the 'black box' problem, as their internal workings and algorithmic implementations are often hidden from users. This can create a dependency where reviewers are making crucial decisions based on tools they neither fully understand nor control. Similarly, the efficiency of automated tools might lead reviewers to overly depend on them, or lead to review projects being undertaken by people who lack the necessary understanding of the systematic review process.

Conclusion

This article has presented some ways in which computational methods can be integrated into systematic review projects to deal with the challenges of increased size and specialisation of scientific corpora. While the uncertainty connected to parsing semantic content algorithmically means that extra care must be taken in the design and implementation phases of a project, it is probable that most review projects in the future will have to integrate ML/AI techniques to plausibly claim that most or all relevant literature has been included and made part of the analysis. Gaining experience with such techniques can help in increasing understanding for how computational text analysis works, and how to ameliorate some of the drawbacks of introducing quantitative text analysis into an analysis practice which relies on meaning and interpretation. There is still much to learn.

Declarations and conflicts of interest

Research ethics statement

Not applicable to this article.

Consent for publication statement

Not applicable to this article.

Conflicts of interest statement

The author declares no conflicts of interest with this work. All efforts to sufficiently anonymise the author during peer review of this article have been made. The author declares no further conflicts with this article.

References

Ayanwale, M.A., Molefi, R.R. and Oyeniran, S. (2024) 'Analyzing the evolution of machine learning integration in educational research: A bibliometric perspective'. *Discover Education*, 3 (1), 47. [CrossRef]

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A.S.C., Ananiadou, S., Liao, J. and Macleod, M.R. (2019) 'Machine learning algorithms for systematic review: Reducing workload in a preclinical review of animal studies and reducing human screening error'. *Systematic Reviews*, 8 (1), 23. [CrossRef] [PubMed]

Bearman, M., Smith, C.D., Carbone, A., Slade, S., Baik, C., Hughes-Warrington, M. and Neumann, D.L. (2012) 'Systematic review methodology in higher education'. *Higher Education Research & Development*, 31 (5), 625–40. [CrossRef]

Bhatti, F., Mowforth, O., Butler, M., Bhatti, Z., Fard, A.R., Kuhn, I. and Davies, B.M. (2023) 'Meeting the shared goals of a student-selected component: Pilot evaluation of a collaborative systematic review'. *JMIR Medical Education*, 9 (1), e39210. [CrossRef] [PubMed]

Bianchini, F., Calamo, M., De Luzi, F., Macrì, M. and Mecella, M. (2024) 'Enhancing complex linguistic tasks resolution through fine-tuning LLMs, RAG and Knowledge Graphs (Short paper)'. In J.P.A. Almeida, C.D. Ciccio and C. Kalloniatis (eds), *Advanced Information Systems Engineering Workshops*. Cham: Springer Nature Switzerland, 147–55. [CrossRef]

Bird, J.J., Ekárt, A. and Faria, D.R. (2023) 'Chatbot interaction with artificial intelligence: Human data augmentation with T5 and language transformer ensemble for text classification'. *Journal of Ambient Intelligence and Humanized Computing*, 14 (4), 3129–44. [CrossRef]

Bornmann, L., Haunschild, R. and Mutz, R. (2021) 'Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases'. *Humanities and Social Sciences Communications*, 8 (1), 1–15. [CrossRef]

Buhagiar, K. and Anand, A. (2023) 'Synergistic triad of crisis management: Leadership, knowledge management and organizational learning'. *International Journal of Organizational Analysis*, 31 (2), 412–29. [CrossRef]

Campbell, F., Tricco, A.C., Munn, Z., Pollock, D., Saran, A., Sutton, A., White, H. and Khalil, H. (2023) 'Mapping reviews, scoping reviews, and evidence and gap maps (EGMs): The same but different – the "Big Picture" review family'. *Systematic Reviews*, 12 (1), 45. [CrossRef]

Campbell, M., McKenzie, J.E., Sowden, A., Katikireddi, S.V., Brennan, S.E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Thomas, J., Welch, V. and Thomson, H. (2020) 'Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline'. *BMJ*, 368: l6890. [CrossRef]

Chandrasekaran, D. and Mago, V. (2022) 'Evolution of semantic similarity – A survey'. *ACM Computing Surveys*, 54 (2), 1–37. [CrossRef]

Cian, H. (2021) 'Sashaying across party lines: Evidence of and arguments for the use of validity evidence in qualitative education research'. *Review of Research in Education*, 45 (1), 253–90. [CrossRef]

Coe, R., Waring, M., Hedges, L.V. and Ashley, L.D. (2021) *Research Methods and Methodologies in Education*. Los Angeles: Sage.

De Bellis, N. (2009) *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Lanham, MD: Scarecrow Press.

de la Torre-López, J., Ramírez, A. and Romero, J.R. (2023) 'Artificial intelligence to automate the systematic review of scientific literature'. *Computing*, 105 (10), 2171–94. [CrossRef]

El-Kassas, W.S., Salama, C.R., Rafea, A.A. and Mohamed, H.K. (2021) 'Automatic text summarization: A comprehensive survey'. *Expert Systems with Applications*, 165, 113679. [CrossRef]

Gamieldien, Y. (2023) 'Innovating the Study of Self-regulated Learning: An exploration through NLP, generative AI, and LLMs'. PhD thesis, Virginia Polytechnic Institute and State University, Blacksburg, USA. Accessed 5 November 2024. http://hdl.handle.net/10919/116274.

Ghosal, A., Nandy, A., Das, A.K., Goswami, S. and Panday, M. (2020) 'A short review on different clustering techniques and their applications'. In J.K. Mandal and D. Bhattacharya (eds), *Emerging Technology in Modelling and Graphics*. Singapore: Springer, 69–83 [CrossRef]

Gkevrou, M. and Stamovlasis, D. (2022) 'Illustration of a software-aided content analysis methodology applied to educational research'. *Education Sciences*, 12 (5), 328. [CrossRef]

Glänzel, W. and Debackere, K. (2022) 'Various aspects of interdisciplinarity in research and how to quantify and measure those'. *Scientometrics*, 127 (9), 5551–69. [CrossRef]

Gorbea Díaz, L., Chopel, A., Fernós Sagebién, A., Bonilla Marrero, L., Rivera Figueroa, G., Pecci Zegrí, N., Cardona, A., Mulero Oliveras, J., La Santa, L. and Sánchez Rey, P. (2023) 'Collecting and communicating perishable data in a post-disaster context: Rapid research and rapid dissemination'. *Frontiers in Sociology*, 8, 959765. [CrossRef]

Gusenbauer, M. and Haddaway, N.R. (2020) 'Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources'. *Research Synthesis Methods*, 11 (2), 181–217. [CrossRef]

Harari, M.B., Parola, H.R., Hartwell, C.J. and Riegelman, A. (2020) 'Literature searches in systematic reviews and meta-analyses: A review, evaluation, and recommendations'. *Journal of Vocational Behavior*, 118, 103377. [CrossRef]

Held, M., Laudel, G. and Gläser, J. (2021) 'Challenges to the validity of topic reconstruction'. *Scientometrics*, 126 (5), 4511–36. [CrossRef]

Hewage, N. and Meedeniya, D. (2022) 'Machine learning operations: A survey on MLOps tool support'. arXiv, 2202.10169. [CrossRef]

Higham, K., Contisciani, M. and De Bacco, C. (2022) 'Multilayer patent citation networks: A comprehensive analytical framework for studying explicit technological relationships'. *Technological Forecasting and Social Change*, 179, 121628. [CrossRef]

Huang, C., Yang, C., Wang, S., Wu, W., Su, J. and Liang, C. (2020) 'Evolution of topics in education research: A systematic review using bibliometric analysis'. *Educational Review*, 72 (3), 281–97. [CrossRef]

Ioannidis, J.P.A. (2022) 'Systematic reviews for basic scientists: A different beast'. *Physiological Reviews*, 103 (1), 1–5. [CrossRef] [PubMed]

Jackson, M., McTier, L., Brooks, L.A. and Wynne, R. (2022) 'The impact of design elements on undergraduate nursing students' educational outcomes in simulation education: Protocol for a systematic review'. *Systematic Reviews*, 11 (1), 52. [CrossRef] [PubMed]

Jethani, N., Jones, S., Genes, N., Major, V.J., Jaffe, I.S., Cardillo, A.B., Heilenbach, N., Ali, N.F., Bonanni, L.J., Clayburn, A.J., Khera, Z., Sadler, E.C., Prasad, J., Schlacter, J., Liu, K., Silva, B., Montgomery, S., Kim, E.J., Lester, J., Hill, T.M., Avoricani, A., Chervonski, E., Davydov, J., Small, W., Chakravartty, E., Grover, H., Dodson, J., Brody, A.A., Aphinyanaphongs, Y. and Razavian, N. (2023) 'Evaluating ChatGPT in information extraction: A case study of extracting cognitive exam dates and scores'. medRxiv. [CrossRef]

Jiang, X. and Liu, J. (2023) 'Extracting the evolutionary backbone of scientific domains: The semantic main path network analysis approach based on citation context analysis'. *Journal of the Association for Information Science and Technology*, 74 (5), 546–69. [CrossRef]

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J. and Kasneci, G. (2023) 'ChatGPT for good? On opportunities and challenges of large language models for education'. *Learning and Individual Differences*, 103, 102274. [CrossRef]

Keraghel, I., Morbieu, S. and Nadif, M. (2024) 'Beyond words: A comparative analysis of LLM embeddings for effective clustering'. In I. Miliou, N. Piatkowski and P. Papapetrou (eds), *Advances in Intelligent Data Analysis XXII*. Cham: Springer Nature Switzerland, 205–16. [CrossRef]

Khurana, D., Koli, A., Khatter, K. and Singh, S. (2023) 'Natural language processing: State of the art, current trends and challenges'. *Multimedia Tools and Applications*, 82 (3), 3713–44. [CrossRef]

Kossmeier, M., Tran, U.S. and Voracek, M. (2020) 'Charting the landscape of graphical displays for meta-analysis and systematic reviews: A comprehensive review, taxonomy, and feature analysis'. *BMC Medical Research Methodology*, 20 (1), 26. [CrossRef]

Kusa, W., Zuccon, G., Knoth, P. and Hanbury, A. (2023) 'Outcome-based evaluation of systematic review automation'. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*. ICTIR '23. New York: Association for Computing Machinery, 125–33. [CrossRef]

Kuzi, S., Zhang, M., Li, C., Bendersky, M. and Najork, M. (2020) 'Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach'. arXiv, 2010.01195. [CrossRef]

Lauriola, I., Lavelli, A. and Aiolli, F. (2022) 'An introduction to deep learning in natural language processing: Models, techniques, and tools'. *Neurocomputing*, 470, 443–56. [CrossRef]

Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M.-I., Noel-Storr, A., Rader, T., Shokraneh, F., Thomas, J. and Wieland, L.S., on behalf of the Cochrane Information Retrieval Methods Group (2019) 'Searching for and selecting studies'. In J.P.T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M.J. Page and V.A. Welch (eds), *Cochrane Handbook for Systematic Reviews of Interventions*. New York: John Wiley & Sons, 67–107. [CrossRef]

Liu, S., Wu, J., Bao, J., Wang, W., Hovakimyan, N. and Healey, C.G. (2024) 'Towards a robust retrieval-based summarization system'. arXiv, 2403.19889. [CrossRef]

Liu, Z., He, X., Liu, L., Liu, T. and Zhai, X. (2023) 'Context matters: A strategy to pre-train language model for science education'. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda and O.C. Santos (eds), *Artificial Intelligence in Education: Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky*. Cham: Springer Nature Switzerland, 666–74. [CrossRef]

Marshall, I.J. and Wallace, B.C. (2019) 'Toward systematic review automation: A practical guide to using machine learning tools in research synthesis'. *Systematic Reviews*, 8 (1), 163. [CrossRef] [PubMed]

Min, B., Ross, H., Sulem, E., Pouran Ben Veyseh, A., Nguyen, T.H., Sainz, O., Agirre, E., Heinz, I. and Roth, D. (2021) 'Recent advances in natural language processing via large pre-trained language models: A survey'. arXiv, 2111.01243. [CrossRef]

Newman, M. and Gough, D. (2020) 'Systematic reviews in educational research: Methodology, perspectives and application'. In O. Zawacki-Richter, M. Kerres, S. Bedenlier, M. Bond and K. Buntins (eds), *Systematic Reviews in Educational Research: Methodology, perspectives and application*. Wiesbaden: Springer Fachmedien, 3–22. [CrossRef]

Occhipinti, A., Rogers, L. and Angione, C. (2022) 'A pipeline and comparative study of 12 machine learning models for text classification'. *Expert Systems with Applications*, 201, 117193. [CrossRef]

O'Connor, A.M., Tsafnat, G., Thomas, J., Glasziou, P., Gilbert, S.B. and Hutton, B. (2019) 'A question of trust: Can we build an evidence base to gain trust in systematic review automation technologies?' *Systematic Reviews*, 8 (1), 143. [CrossRef]

Pak, K. and Ravitch, S.M. (2021) *Critical Leadership Praxis for Educational and Social Change*. New York: Teachers College Press.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J. and Wu, X. (2024) 'Unifying large language models and knowledge graphs: A roadmap'. *IEEE Transactions on Knowledge and Data Engineering*, 36 (7), 3580–99. [CrossRef]

Park, S., Wang, A.Y., Kawas, B., Liao, Q.V., Piorkowski, D. and Danilevsky, M. (2021) 'Facilitating knowledge sharing from domain experts to data scientists for building NLP models'. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. IUI '21. New York: Association for Computing Machinery, 585–96. [CrossRef]

Polanin, J.R., Maynard, B.R. and Dell, N.A. (2017) 'Overviews in education research: A systematic review and analysis'. *Review of Educational Research*, 87 (1), 172–203. [CrossRef]

Polat, M. (2022) 'Exploring educational research during the COVID-19 pandemic: 2020–2021'. *FIRE: Forum for International Research in Education*, 7 (2), 86–104. [CrossRef]

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. and Manning, C.D. (2020) 'Stanza: A Python natural language processing toolkit for many human languages'. arXiv, 2003.07082. [CrossRef]

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J. (2020) 'Exploring the limits of transfer learning with a unified text-to-text transformer'. *Journal of Machine Learning Research*, 21 (140), 1–67. [CrossRef]

Rethlefsen, M.L., Kirtley, S., Waffenschmidt, S., Ayala, A.P., Moher, D., Page, M.J., Koffel, J.B. and PRISMA-S Group (2021) 'PRISMA-S: An extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews'. *Systematic Reviews*, 10 (1), 39. [CrossRef]

Rungta, M., Singh, J., Mohammad, S.M. and Yang, D. (2022) 'Geographic citation gaps in NLP research'. arXiv, 2210.14424. [CrossRef]

Shahjahan, R.A., Estera, A.L., Surla, K.L. and Edwards, K.T. (2022) '"Decolonizing" curriculum and pedagogy: A comparative review across disciplines and global higher education contexts'. *Review of Educational Research*, 92 (1), 73–113. [CrossRef]

Sjögårde, P. and Ahlgren, P. (2020) 'Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties'. *Quantitative Science Studies*, 1 (1), 207–38. [CrossRef]

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S. and Boomgaarden, H.G. (2020) 'In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis'. *Political Communication*, 37 (4), 550–72. [CrossRef]

Susnjak, T., Hwang, P., Reyes, N.H., Barczak, A.L.C., McIntosh, T.R. and Ranathunga, S. (2024) 'Automating research synthesis with domain-specific large language model fine-tuning'. arXiv, 2404.08680. [CrossRef]

Tao, J., Zhou, L. and Hickey, K. (2022) 'Making sense of the black-boxes: Toward interpretable text classification using deep learning models'. *Journal of the Association for Information Science and Technology*, 74 (6), 685–700. [CrossRef]

Thelwall, M. and Sud, P. (2022) 'Scopus 1900–2020: Growth in articles, abstracts, countries, fields, and journals'. *Quantitative Science Studies*, 3 (1), 37–50. [CrossRef]

Thomas, D.A. (2014) 'Searching for significance in unstructured data: Text mining with Leximancer'. *European Educational Research Journal*, 13 (2), 235–56. [CrossRef]

Tong, A., Flemming, K., McInnes, E., Oliver, S. and Craig, J. (2012) 'Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ'. *BMC Medical Research Methodology*, 12 (1), 181. [CrossRef] [PubMed]

Van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L. and Oberski, D.L. (2021) 'An open source machine learning framework for efficient and transparent systematic reviews'. *Nature Machine Intelligence*, 3 (2), 125–33. [CrossRef]

Van Dinter, R., Tekinerdogan, B. and Catal, C. (2021) 'Automation of systematic literature reviews: A systematic literature review'. *Information and Software Technology*, 136, 106589. [CrossRef]

Vayansky, I. and Kumar, S.A.P. (2020) 'A review of topic modeling methods'. *Information Systems*, 94, 101582. [CrossRef]

Vital, A. and Amancio, D.R. (2022) 'A comparative analysis of local similarity metrics and machine learning approaches: Application to link prediction in author citation networks'. *Scientometrics*, 127 (10), 6011–28. [CrossRef]

Wagner, G., Lukyanenko, R. and Paré, G. (2022) 'Artificial intelligence and the conduct of literature reviews'. *Journal of Information Technology*, 37 (2), 209–26. [CrossRef]

Wang, Q. and Waltman, L. (2016) 'Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus'. *Journal of Informetrics*, 10 (2), 347–64. [CrossRef]

Wang, S., Mao, J., Cao, Y. and Li, G. (2022) 'Integrated knowledge content in an interdisciplinary field: Identification, classification, and application'. *Scientometrics*, 127 (11), 6581–614. [CrossRef]

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P.S. and Wen, Q. (2024) 'Large language models for education: A survey and outlook'. arXiv, 2403.18105. [CrossRef]

Wang, W.-T. and Wu, S.-Y. (2021) 'Knowledge management based on information technology in response to COVID-19 crisis'. *Knowledge Management Research & Practice*, 19 (4), 468–74. [CrossRef]

Wankhade, M., Rao, A.C.S. and Kulkarni, C. (2022) 'A survey on sentiment analysis methods, applications, and challenges'. *Artificial Intelligence Review*, 55 (7), 5731–80. [CrossRef]

Weisser, T., Sassmannhausen, T., Ohrndorf, D., Burggräf, P. and Wagner, J. (2020) 'A clustering approach for topic filtering within systematic literature reviews'. *MethodsX*, 7, 100831. [CrossRef] [PubMed]

Xie, Q., Bishop, J.A., Tiwari, P. and Ananiadou, S. (2022) 'Pre-trained language models with domain knowledge for biomedical extractive summarization'. *Knowledge-Based Systems*, 252, 109460. [CrossRef]

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y. and Gašević, D. (2024) 'Practical and ethical challenges of large language models in education: A systematic scoping review'. *British Journal of Educational Technology*, 55 (1), 90–112. [CrossRef]