Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 9 No. 3 (2021) 1-12
ISSN 2295-3159

Staying at the front line of literature: How can topic modelling help researchers follow recent studies?

Joni Lämsä[1], Catalina Espinoza[2], Ari Tuhkala[3], & Raija Hämäläinen[1]

[1]Department of Education, University of Jyväskylä, Finland
[2]Center for Advanced Research in Education, University of Chile, Chile
[3]Finnish Institute for Educational Research, University of Jyväskylä, Finland

Article received 23 June 2020 / Article revised 20 December / Accepted 26 March 2021 / Available online 14 April

Abstract

Staying at the front line in learning research is challenging because many fields are rapidly developing. One such field is research on the temporal aspects of computer-supported collaborative learning (CSCL). To obtain an overview of these fields, systematic literature reviews can capture patterns of existing research. However, conducting systematic literature reviews is time-consuming and do not reveal future developments in the field. This study proposes a machine learning method based on topic modelling that takes articles from a systematic literature review on the temporal aspects of CSCL (49 original articles published before 2019) as a starting point to describe the most recent development in this field (52 new articles published between 2019 and 2020). We aimed to explore how to identify new relevant articles in this field and relate the original articles to the new articles. First, we trained the topic model with the Results, Discussion, and Conclusion sections of the original articles, enabling us to correctly identify 74% (n = 17) of new and relevant articles. Second, clusterisation of the original and new articles indicated that the field has advanced in its new and relevant articles because the topics concerning the regulation of learning and collaborative knowledge construction related 26 original articles to 10 new articles. New irrelevant studies typically emerged in clusters that did not include any specific topic with a high topic occurrence. Our method may provide researchers with resources to follow the patterns in their fields instead of conducting repetitive systematic literature reviews.

Keywords: automatic content analysis; computer-supported collaborative learning; literature review; temporal analysis; topic model

Info corresponding author: Joni Lämsä, P.O. Box 35, FI-40014, University of Jyväskylä, Finland, joni.lamsa@jyu.fi
DOI: *https://doi.org/10.14786/flr.v9i3.645*

1. Introduction

Research in learning sciences has become more interdisciplinary because increasingly complex datasets and methods may require the expertise of computer scientists and signal processors. This interdisciplinary collaboration opens up the possibility of new publication forums in the learning sciences. However, this could also make thorough systematic or thematic literature reviews (see Gruber et al., 2020) even more arduous. Thus, it would be useful if the vast amount of work done by scholars when conducting systematic literature reviews could be exploited when monitoring how a specific line of research would proceed. If relevant future studies can be automatically identified and related to previous research, this would decrease the need to perform recurring systematic literature reviews on similar topics, thus affording researchers more working hours to advance in their fields. To address these aspirations, we present a machine learning–based method that takes articles from a manual systematic literature review as a starting point to describe the recent developments in the field. We illustrate the potential of our innovative method in the context of research focusing on the temporal analysis of computer-supported collaborative learning (CSCL). This field of research forms a particularly promising basis for studying its progress because the studies focusing on the temporal aspects of CSCL are increasingly being published and involve interdisciplinary collaboration (e.g., Hadwin, 2021; Lämsä et al., 2021).

In this study, we define the temporal analysis of CSCL as analysing the characteristics of events or the interrelations between these events over time. The events may relate to learner interaction, thoughts and ideas developed during the interaction and the use of technological resources to mediate the interaction (see Lämsä et al., 2021). A temporal analysis of CSCL may benefit both practitioners and researchers by revealing *how* (not only *what*) learning occurs in CSCL settings (Lämsä, 2020), particularly now when COVID-19 highlights the need for effective CSCL more than ever (Järvelä & Rosé, 2020). When we manually reviewed the literature focusing on the temporal aspects of CSCL (see section 2 and Lämsä et al., 2021), we found that the interdisciplinary collaboration in this field has caused challenges regarding the commensurability and comparability of the studies. Particularly, the studies seemed to be fragmented in terms of their theoretical frameworks (cf. Hew et al., 2019), methodologies, and results and implications. This finding implies that both practitioners and researchers may struggle with staying at the front line concerning the big picture of CSCL and its research because of this fragmentation.

Practitioners may benefit from our method if it can filter applicable research to support them in the design and implementation of research-based CSCL innovations. Similarly, our method can benefit researchers because it can illustrate whether and how the recent research has contributed to prior studies. We investigate the added value of our method for practitioners and researchers by addressing the following research questions:

RQ1: How and to what extent can a machine learning–based method be used to identify new relevant articles in the field of manual systematic literature review?

RQ2: How and to what extent can the machine learning–based method be used to relate new and original articles to each other?

2. Methodology

When manually reviewing the literature on the temporal aspects of CSCL in February 2019 (see Lämsä et al., 2021), we carefully selected the search terms concerning temporality, collaborative learning, and computer-supported learning. We used the Education Resources Information Center (ERIC), Scopus, and Web of Science databases and identified 436 articles, of which we manually screened and assessed their eligibility. In this study, we included 49 peer-reviewed journal articles that focused on the temporal analysis of CSCL for further analysis (*original articles*). To find new articles,

we repeated the literature searches with the same search terms and databases in February 2020 as for the original articles. The searches found 88 articles that had been published between February 2019 and 2020. From these 88 articles, we excluded 36 articles, of which 31 were duplicates, three had no full text available, one was a conference proceeding article, and one was already included in the set of the original articles. In the following analyses, we refer to these included 52 peer-reviewed journal articles as a set of *new articles*.

The utilised machine learning–based method was grounded on a natural language processing technique known as topic modelling, which is based on statistical algorithms that find topics in a collection of documents (Boyd-Graber et al., 2017). These topics are ranked lists of words, where each word has a probability of belonging to a topic (see Table 1), or more formally, topics are probability distributions over vocabularies. In the following sections, we describe how the original articles were exploited to build the topic models that, in turn, were used to identify the new relevant articles (RQ1) and relate them to the original articles (RQ2). Figure 1 summarises our procedure.

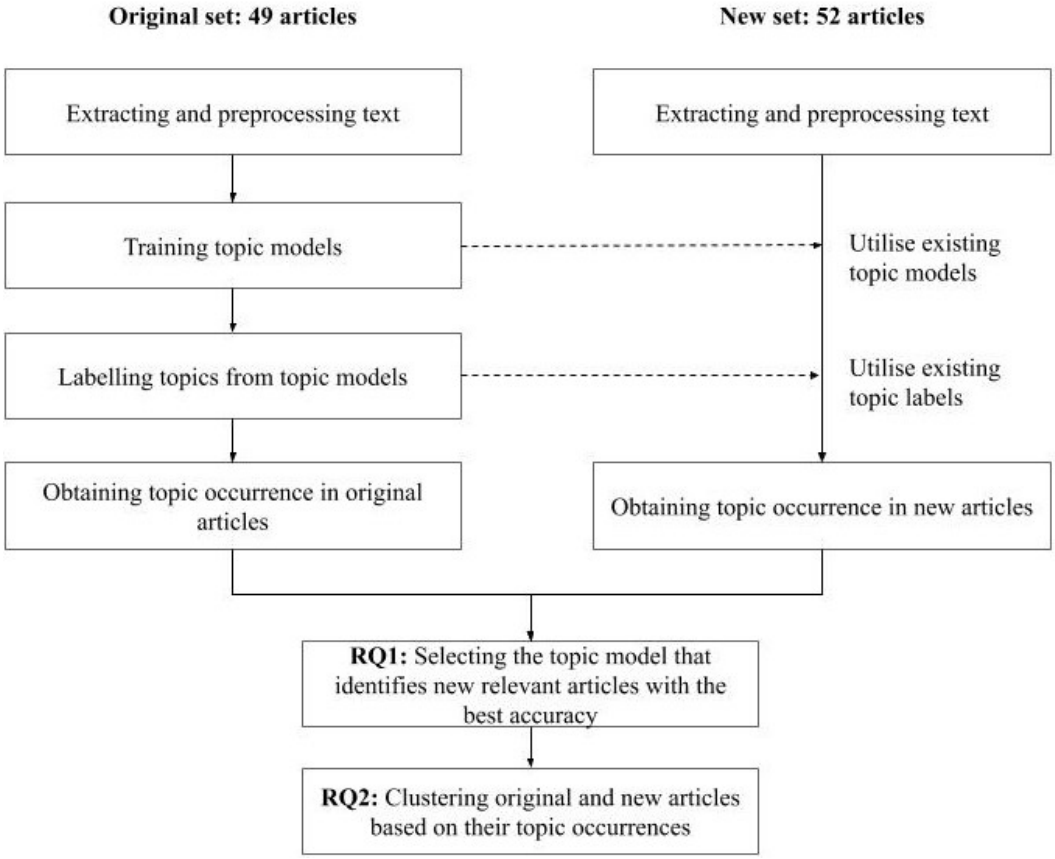


Figure 1: Procedure for describing original and new articles to address the research questions (RQs)

2.1 Extracting and preprocessing text

First, we extracted raw text from the original and new articles and removed tables, figures, formulas, bullet points, footnotes, and page numbers. Second, we separated the different sections of the articles under the following headings: Introduction, Theoretical Framework, Methodology, Results, Discussion, and Conclusion. However, because not all the articles had all of these sections (e.g., an article may have a combined Results and Discussion section), we decided to combine the sections into three wider sections: (1) Introduction and Theoretical Framework, (2) Methodology, and (3) Results, Discussion, and Conclusion. Then, we carried out text preprocessing, including common text cleaning, such as transforming text to lowercase and removing symbols and infrequent words. Finally, we utilised

the Natural Language Toolkit (Bird et al., 2009) to perform word stemming (reducing words to their root form) and common English stop word removal (e.g., the, at, is).

2.2 Training topic models

We used latent Dirichlet allocation (LDA) (Blei et al., 2003) and the Gensim library (Rehurek & Sohjka, 2010) to train topic models for each section 1–3 of the original articles. The output of training topic models includes both a list of topics and the trained topic model itself. The trained topic model can process new text and measure the presence of the listed topics. We performed a sensitivity analysis based on topic coherence values (provided by the Gensim library) to find an appropriate number of topics for each section. As an outcome, we had trained three topic models, one for each section, that all included 17 topics.

2.3 Labelling topics from topic models

The trained topic models contained a list of topics found in each section. We labelled the topics by analysing the most representative words and utilising expert knowledge from the manual systematic review of the literature. If possible, we labelled the topics based on the theoretical framework to which the most representative words refer. We demonstrate this idea in Table 1 using topic models for section 3 as an example, presenting the labels and the 10 most representative words. For example, topic 1 (temporal aspects of CSCL) is a generic topic that illustrates a stage in the temporal analysis procedure. Namely, researchers *code messages* of *groups* of *students,* after which they *analyse* the typical *sequences* of *messages*. This kind of *sequential analysis* reveals what kind of *messages* follow each other in a short temporal context whose duration may be a few *messages* (the words with italics refer to the 10 most representative words from topic 1).

2.4 Obtaining topic occurrence in original articles

In LDA, articles are represented as lists of topic probabilities; the goal is to find the topic probabilities of a document that are better suited to rebuild the document by randomly selecting words. For example, if an article has a higher topic probability for topic 16 compared with other topics (see Table 1), most of the words in the article can be selected from the top of topic 16. We refer to topic probabilities in an article as a topic occurrence to distinguish them from words' probabilities inside a topic. When we used topic models for sections 1–3, we obtained 51 topic occurrences for each original article (17 topic occurrences for each topic model).

2.5 Obtaining topic occurrence in new articles

The process used for the new articles was very similar to the one applied to the old articles (Figure 1). The only difference was that we directly applied the trained topic models for sections 1–3 to obtain the topic occurrences of the new articles. We illustrate the topic occurrences of original, new relevant, and new irrelevant articles using the topic model for section 3 in Figure 2. For each article, some topics have a higher probability than the rest (e.g., topic 16 is more relevant to an original article than to a new irrelevant article; see (a) and (c) in Figure 2). Therefore, we expect to find semantic similarity between topic occurrences that have shorter distances.

Table 1

Five topics and the assigned topic labels, including the 10 most representative words from the topic model for section 3 (Results, Discussion, and Conclusion).

Topic 1: Temporal aspects of computer-supported collaborative learning	**Topic 7: Regulation of learning and learning performance**	**Topic 8: Regulation of learning**	**Topic 11: Socially shared metacognitive regulation (SSMR)**	**Topic 16: Collaborative knowledge construction**
NUMBER	Model	Regul	SSMR	Discuss
Student	Group	Learn	Process	Group
Signific	Perform	Collabor	Phase	Student
Code	Focus	Task	Thread	Knowledg
Group	Student	Social	Studi	Behaviour
Analysi	SSRL[1]	Share	Research	Construct
Show	Collabor	Student	Differ	Learn
Sequenc	Challeng	Group	Inquiri	Result
Knowledg	Differ	Result	Note	Process
Messag	Individu	Discuss	Data	Pattern

[1]Socially shared regulation of learning

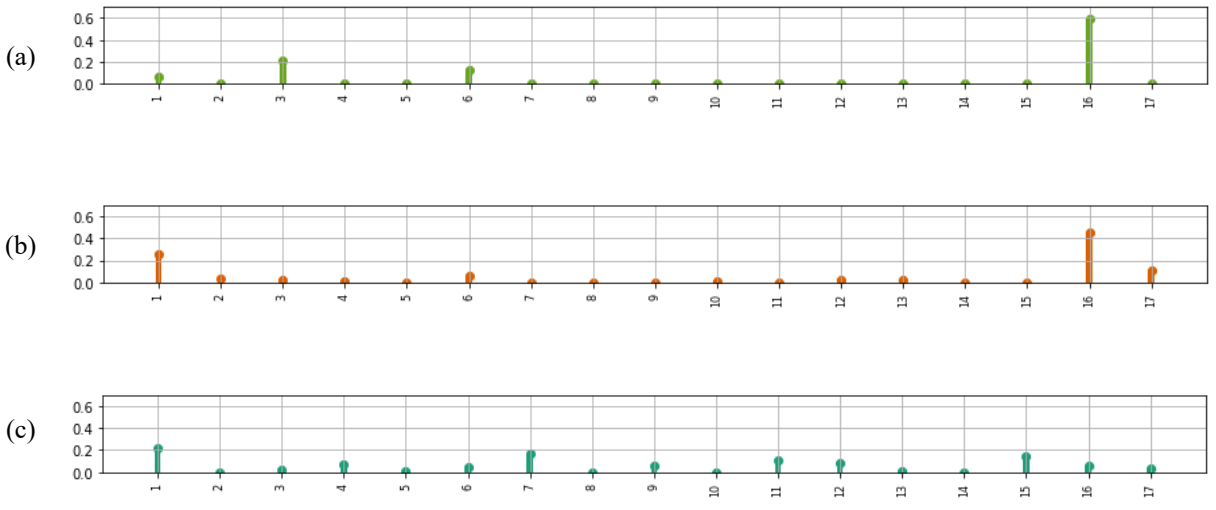


Figure 2: The topic occurrence of (a) an original article, (b) a new relevant article, and (c) a new irrelevant article obtained using the topic model for section 3. The distance between (a) and (b) was 0.33, between (a) and (c) was 0.65, and between (b) and (c) was 0.49.

To answer RQ1, the first and second authors screened and labelled the new 52 articles manually as relevant or irrelevant regarding the analysis of the temporal aspects of CSCL. In the first phase, we screened the journal and title of the articles and labelled the studies that did not have a learning or

instructional context as irrelevant (n = 27, e.g., studies from environmental sciences). In the second phase, we also screened the abstract of the articles and labelled the studies that did not focus on CSCL and analyse its temporal aspects as irrelevant (n = 2, e.g., a study that focused merely on learning performance). We solved the disagreements between the first and second authors in the common meetings among all the authors. Altogether, 23 new articles focused on the analysis of the temporal aspects of CSCL (relevant), while 29 articles did not (irrelevant). Next, for each topic model, we measured the distance between the corresponding topic occurrences of the new articles and original articles. The shorter the distance between two articles, the more similar the topic occurrences (Figure 2). For each new article, we kept the distance to the closest original article (i.e., the most similar because a new relevant article might not be related to every article in the manual systematic literature review). Finally, we compared the distances between the relevant and irrelevant articles. We selected the most suitable topic model so that the topic occurrences of the new relevant articles were similar to the ones from the original articles.

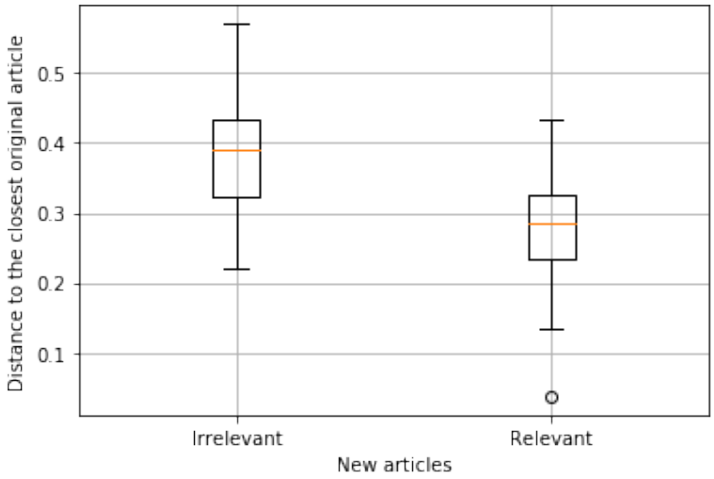
To answer RQ2, we used the articles' topic occurrences from the previously selected topic model (RQ1). We measured the similarity between topic occurrences using the Euclidean distance, and we applied hierarchical clustering to find groups of similar articles. We performed the clustering in three levels: the root, two subgroups, and the leaves. The root of the clustering contains all the articles: 52 new articles and 49 original articles. The root was then divided into two subgroups, denoting the greatest distance between the articles belonging to different subgroups. The leaves are groups of articles of varying sizes. We interpreted the clusters by examining the topic occurrences (Figure 2) and previously assigned topic labels (Table 1).

3. Results

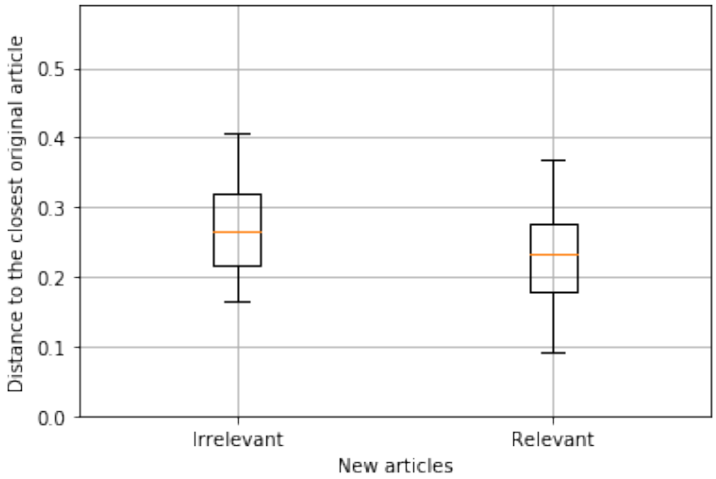
3.1 The topic model trained with the Results, Discussion, and Conclusion sections identified new relevant articles most accurately.

We identified new relevant articles relating to the temporal aspects of CSCL by measuring the distance between a new article and the closest original article. The results showed that for the three topic models, the relevant new articles were closer to the original articles than the irrelevant articles (Figure 3). Particularly, the topic model for section 3, which we trained with Results, Discussion, and Conclusion sections, gave the best results because the distance between the new relevant articles and the closest original article overlapped the least with the distance between new irrelevant articles and the closest original article [Figure 3 (c)]. When we used the topic model for section 3 and the distance of 0.27 as a threshold, 71% of the new articles, which were closer than the threshold, were relevant. Those relevant articles represent 74% of the total relevant articles, which minimised the number of irrelevant articles. When we used topic models for sections 1 and 2, the distances between the original articles and new relevant articles overlapped more with new irrelevant articles [Figure 3 (a) and (b)]. Table 2 summarises our results if the distance of 0.27 is considered for the threshold.

(a)



(b)



(c)

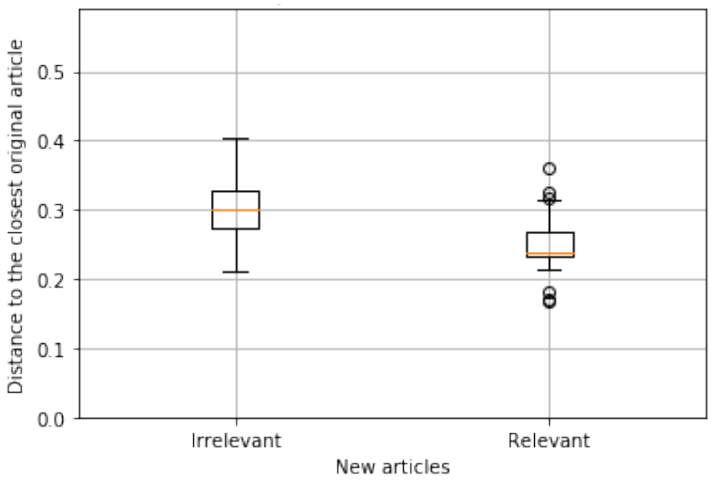


Figure 3: Boxplots of the distances between relevant and irrelevant new articles and the closest original article separately for (a) topic model for section 1, (b) topic model for section 2, and (c) topic model for section 3.

Table 2

The numbers of relevant and irrelevant articles identified and missed when using three topic models

	Topic model (section 1): Introduction and Theoretical framework	Topic model (section 2): Methodology	Topic model (section 3): Results, Discussion, and Conclusion
Relevant identified (true positives)	10	17	17
Relevant missed (false negatives)	13	6	6
Irrelevant identified as relevant (false positive)	3	14	7
Irrelevant identified as irrelevant (true negatives)	26	15	22
Precision (proportion of the true positives to the sum of true and false positives)	0.77	0.55	0.71
Recall (proportion of the true positives to the sum of the true positives and false negatives)	0.43	0.74	0.74

3.2 A few topics with high topic occurrence relate new relevant to original articles

Figure 4 shows the outcome of the hierarchical clustering. When interpreting Figure 4, based on the CSCL theoretical frameworks, a few topics concerning collaborative knowledge construction and regulation of learning relate new relevant articles to original articles. Topic 16 (see Table 1) relates five new relevant articles to 17 original articles (the leaves with double borders), and these articles mostly belong to a smaller subgroup. Topics 7, 8, and 11 (see Table 1) relate five new relevant articles to nine original articles (the leaves with bold borders), and these articles belong to a larger subgroup.

Most of the new irrelevant articles (n = 28) were clustered into three different leaves (Figure 4). From this set, 17 articles appeared in the leaves with different topics. Moreover, 11 articles appeared in the leaf that included only one new relevant article and one original article. Most of the original articles (n = 31) had a topic with a value higher than 0.45. In contrast, the clusters formed by various topics contained articles in which the topic occurrence of the most important topic was less than 0.2, meaning that there were no predominant topics. Because topic occurrence is a probability distribution (must sum up to one), topic occurrence is more scattered if no particular topic is more significant [see Figure 2 (c)]; this feature clusters together most of the irrelevant articles, but it also mixes irrelevant articles with relevant articles that have several important topics. In our case, 12 new relevant articles emerged in the leaves with different topics.

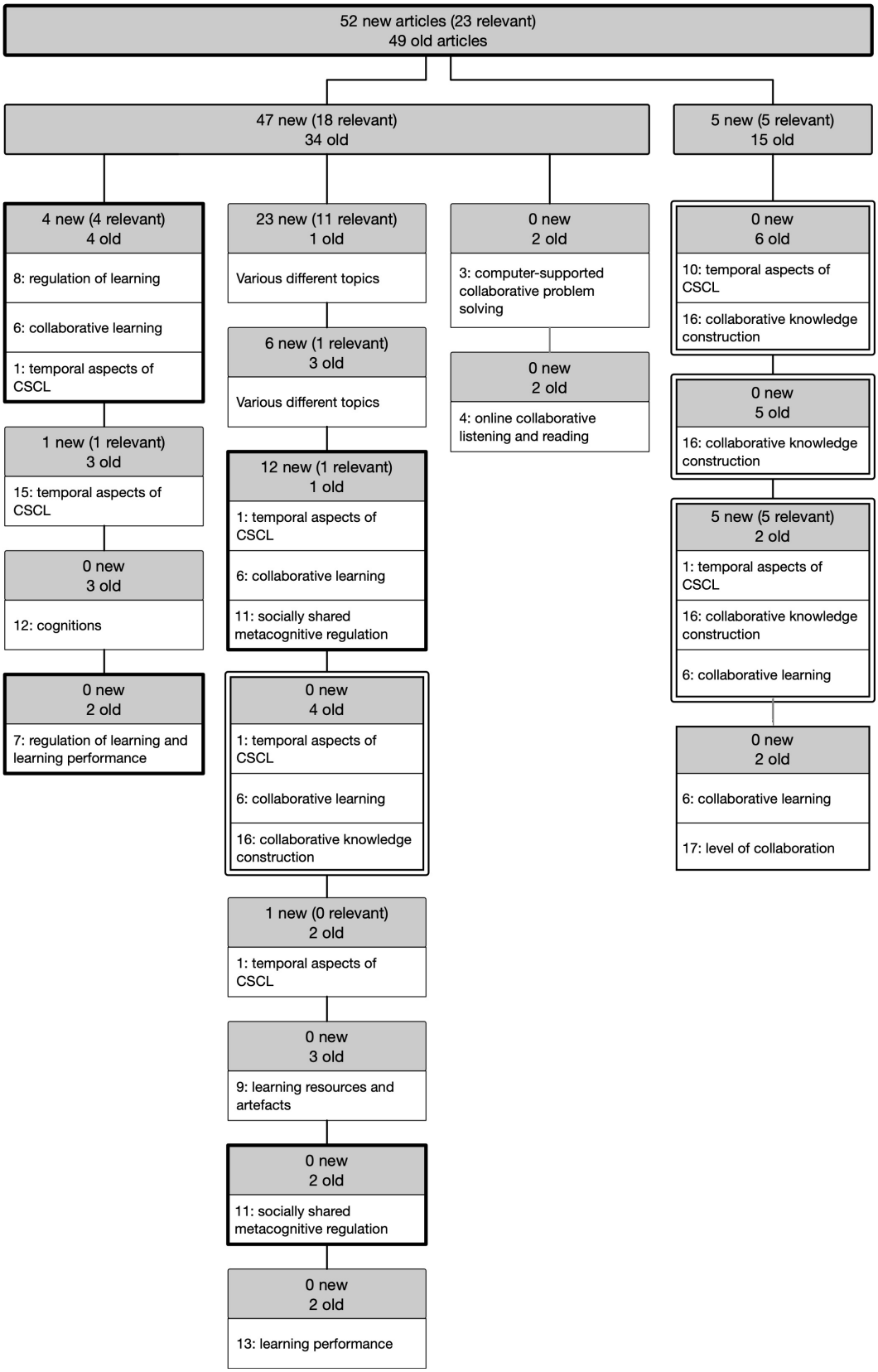


Figure 4: New and original articles' clustered and associated topics. Each leaf has a grey box with the number of new and original articles. The text in the leaves corresponds to the number of the main topics and their labels.

4. Discussion and Conclusion

When considering some of the most-cited journals in the educational research field (*Review of Educational Research* and *Educational Research Review*), systematic literature reviews may 'shape the future of research and practice' (Murphy et al., 2017, p. 2; Alexander, 2020). We showed how a machine learning–based method can be used to identify new relevant articles in the field of the manual systematic literature review (RQ1) and how it can relate new and original articles to each other (RQ2). This novel method may help to follow the evolution of the 'big picture' of the research fields based on multidisciplinary collaboration, such as studies focusing on the temporal analysis of CSCL. Because these studies are published in various forums and involve different theoretical frameworks, methodological approaches, and results and implications, our methodological innovation may reveal *how* literature reviews can 'shape the future of research and practice'.

Even though our method has potential, there are several limitations and critical issues to consider because the current study was an initial attempt to investigate the potential of topic modelling in the context of staying in the front line of literature. First, instead of an 'automatic' method, our method could be called 'semiautomatic' (cf. Tuhkala et al., 2018). For instance, we extracted the texts of different article sections manually. Even though this extraction process could be automatised, we decided to focus on automating more complex phases of our procedure (see Figure 1). In the future, we aim to automatise a pipeline in which all the articles that arise from certain search terms can be processed, filtered according to their relevance, and related to original articles. Second, because the number of original articles was relatively small, we could apply a heuristic to automatically identify the relevant and irrelevant new articles (RQ1; see Table 2). Our heuristic was based on identifying new relevant articles without including too many irrelevant ones (high precision value) or filtering relevant ones (high recall value; see Table 2). In the future, more complex methods can be tested, particularly if there are more articles. Third, we trained the topic model only with original articles (Figure 1), so all the topics can relate to the temporal aspects of CSCL and the analysis of these aspects. Thus, there were no topics that could have properly described new irrelevant articles. We will consider training the topic models by using both original and new articles and using the articles of related systematic literature reviews to capture a broader picture of the field.

Despite these limitations, our innovative method may open up new avenues to follow patterns in the different research fields based on the content of articles, instead of, for example, mere bibliographic information (Chen et al., 2020). A recently published editorial of *Educational Research Review* (Gruber et al., 2020, p. 1) highlighted that systematic literature reviews should 'extend beyond reporting or summarising what has been done in a particular field'. Here, we see our method as more complementary than contradictory to researchers' manual work when using the review approach to address their research questions. Namely, topic modelling (Figure 1) is an unsupervised method, so it may reveal patterns (or topics; see an example in Table 1) from the existing literature to which researchers may not pay attention to. Moreover, our method may assist researchers in some time-consuming tasks when they conduct systematic literature reviews. If considering the preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement (Moher et al., 2009) as an example, machine learning–based methods may help researchers in *identifying* the relevant articles (RQ1) and in *screening* and assessing the *eligibility* of the articles based on their relatedness to the research problems of interest (RQ2). This kind of assistance would allow for investing more resources in a critical review of the *included* articles and, thus, scientifically valuable contributions. At the same time, it is as crucial in machine learning–based methods as it is in manual systematic literature reviews that researchers report their decisions transparently throughout the process (cf. our procedure in Figure 1 and sections 2.1–2.5).

In our research context, the topic model for section 3—which we trained with the Results, Discussion, and Conclusion sections—had the best performance in identifying new relevant articles (RQ1); this may be related to the theoretical fragmentation of studies in the educational technology field (Hew et al., 2019) because original papers had been published in the journals of both computer sciences

and learning sciences. Thus, the topic model for section 1 could not properly separate new relevant and irrelevant articles. Moreover, the methods used to analyse the temporal aspects of CSCL (e.g., sequential analysis) have been used in many disciplines, so the predictive power of the topic model for section 2 might be affected by this issue. In addition to identifying new relevant articles with moderate accuracy by using the topic model for section 3, we could relate new and old articles based on the few topics present in this topic model (RQ2). For example, we found leaves of eight articles (four new relevant and four original articles) and seven articles (five new relevant and two original articles) that seemed to concern the temporal aspects of CSCL in the context of the regulation of learning and collaborative knowledge construction, respectively (Figure 4). These findings may inform both practitioners and researchers by showing widely used theoretical frameworks and providing a 'state of the research' (Murphy et al., 2017, p. 5; see Figure 4).

In the future, when the number of new articles increases, clearer clusters and leaves of original and new articles may emerge. Researchers can follow the fluctuation of the rising research topics by monitoring the size of the leaves (Figure 4). The increasing number of articles would allow for more focused machine learning–based literature reviews so that the procedure for describing new articles (Figure 1) would focus on, for example, a certain theoretical framework through which the temporal aspects of CSCL can be analysed. Our method could also be applied in completely different research fields if there is an existing systematic literature review from that field, and it is possible to train the topic models based on the included articles in the review (section 2.1). As research fields differ from each other and similar fields may have fundamentally different research traditions, further studies could, for example, investigate how to obtain a topic model (section 2.2) and its essential topics (section 2.3) whose topic occurrences (sections 2.4–2.5) could separate studies with fundamentally different epistemological stances. Obtaining topic models and interpreting their essential topics, which researchers can do to address their research aims, require thorough expertise on the research field, in addition to the knowledge and skills to apply machine learning–based methods.

Key points

- Many research fields on learning sciences are developing rapidly, which makes conducting systematic literature reviews a time-consuming task.
- We propose an innovative method that uses an existing systematic literature review to describe the recent developments in the field being reviewed.
- We illustrate the potential of our method using the literature on the temporal analysis of computer-supported collaborative learning.
- Our machine learning–based method identified new relevant articles and related them to the previous literature with moderate accuracy.
- Our method may decrease the need to do recurring systematic literature reviews, giving researchers more working hours to advance their fields.

Acknowledgements

This research was funded by the Academy of Finland [grant numbers 292466 and 318095, the Multidisciplinary Research on Learning and Teaching profiles I and II of University of Jyväskylä].

References

Alexander, P. A. (2020). Methodological guidance paper: The art and science of quality systematic reviews. *Review of Educational Research*, *90*(1), 6–23. https://doi.org/10.3102/0034654319854352

Bird, S., Loper E., & Klein, E. (2009). *Natural language processing with Python*. O'Reilly Media Inc.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boyd-Graber, J. L., Hu, Y., & Mimno, D. (2017). *Applications of topic models* (Vol. 11). Now Publishers Incorporated.

Chen, X., Zou, D., & Xie, H. (2020). Fifty years of *British Journal of Educational Technology*: A topic modeling based bibliometric perspective. *British Journal of Educational Technology*, *51*(3), 692–708. https://doi.org/10.1111/bjet.12907

Gruber, H., Hämäläinen, R. H., Hickey, D. T., Pang, M. F., & Pedaste, M. (2020). Mission and scope of the *Journal Educational Research Review*. *Educational Research Review, 30,* 100328. https://doi.org/10.1016/j.edurev.2020.100328

Hadwin, A. F. (2021). Commentary and future directions: What can multi-modal data reveal about temporal and adaptive processes in self-regulated learning? *Learning and Instruction, 72*, 101287. https://doi.org/10.1016/j.learninstruc.2019.101287

Hew, K. F., Lan, M., Tang, Y., Jia, C., & Lo, C. K. (2019), Where is the 'theory' within the field of educational technology research? *British Journal of Educational Technology*, *50*(3), 956–971. https://doi.org/10.1111/bjet.12770

Järvelä, S., & Rosé, C. P. (2020). Advocating for group interaction in the age of COVID-19. *International Journal of Computer-Supported Collaborative Learning, 15*(2), 143–147. https://doi.org/10.1007/s11412-020-09324-4

Lämsä, J. (2020). *Developing the temporal analysis for computer-supported collaborative learning in the context of scaffolded inquiry* [Doctoral dissertation, University of Jyväskylä]. JYU dissertations, 245. http://urn.fi/URN:ISBN:978-951-39-8248-5

Lämsä, J., Hämäläinen, R., Koskinen, P., Viiri, J., & Lampi, E. (2021). What do we do when we analyse the temporal aspects of computer-supported collaborative learning? A systematic literature review. *Educational Research Review, 33*, 100387. https://doi.org/10.1016/j.edurev.2021.100387

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med, 6*(7), 1–6. https://doi.org/10.1136/bmj.b2535

Murphy, P. K., Knight, S. L., & Dowd, A. C. (2017). Familiar paths and new directions: Inaugural call for manuscripts. *Review of Educational Research, 87*(1), 3–6. https://doi.org/10.3102/0034654317691764

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). ELRA. https://doi.org/10.13140/2.1.2393.1847

Tuhkala, A., Kärkkäinen, T., & Nieminen, P. (2018). Semi-automatic literature mapping of participatory design studies 2006–2016. In *Proceedings of the 15th Participatory Design Conference* (pp. 1–5). Association for Computing Machinery. https://doi.org/10.1145/3210604.3210621