IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 October 2023, accepted 20 October 2023, date of publication 25 October 2023, date of current version 1 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3327624

RESEARCH ARTICLE

Uncovering the Educational Data Mining Landscape and Future Perspective: A Comprehensive Analysis

OZCAN OZYURT[1], HACER OZYURT[1], AND DEEPTI MISHRA[2], (Senior Member, IEEE)
[1]Faculty of Technology, Department of Software Engineering, Karadeniz Technical University, 61080 Trabzon, Turkey
[2]Educational Technology Laboratory, Department of Computer Science (IDI), Norwegian University of Science and Technology, 2815 Gjøvik, Norway

Corresponding author: Deepti Mishra (deepti.mishra@ntnu.no)

This work was supported by the Norwegian University of Science and Technology, Gjøvik, Norway.

ABSTRACT Educational data mining (EDM) enables improving educational systems by using data mining techniques on educational data to analyze students' learning processes to extract valuable information that helps optimize teaching strategies and improve student achievement. EDM has been an important area of research and application in recent years. The aim of this study is to describe the current situation of the EDM field and reveal its future perspective. The study employs descriptive analysis and topic modeling, utilizing a corpus of 2792 studies indexed in the Scopus database since 2007. Firstly, the study determines the document types, distribution by years, prominent authors, countries, subject areas, and journals of the studies in the field of EDM. Then, using topic modeling analysis, which is an unsupervised machine learning technique, the study determines hidden patterns, research interests, and trends within the field. This study is innovative and the first as it reveals latent research interests and trends in the field of EDM through machine learning-based topic modeling-based analysis. The descriptive characteristics of the study emphasize the continuous development of the field and its multidisciplinary aspect. The outputs of the topic modeling analysis reveal that the studies can be grouped into twelve topics. The most frequently studied topic is ''Learning pattern and behavior'', and the topic whose frequency of study increases the most over time is ''Dropout risk prediction''. When comparing the frequency of study of the topics over time to other topics, the first topic that stands out is ''Performance prediction''. The results of this study can be expected to make significant contributions to the field in terms of revealing the big picture of the current literature in the field of EDM and providing a future perspective. Therefore, the results of the study are expected to give direction to the field and provide important insights or guidance to decision makers and education policy makers.

INDEX TERMS Educational data mining, topic modeling, research trends, machine learning.

I. INTRODUCTION

The development of educational technologies has brought about changes in educational processes. The inclusion of internet technologies in educational processes, the diversification of resources and the use of educational software, in short, technology-enhanced learning, created large data pools where data about students are stored [1], [2]. These educational data pools, which are increasing day by day, are a gold mine for education stakeholders [3], [4]. The information that can be discovered in such pools can be used not only to model the learning process, but also to evaluate learning systems and improve the quality of managerial decisions [1]. Data mining or knowledge discovery from databases is defined as the automatic extraction of important patterns from such repositories [5]. In the field of education, institutions and learning environments generate daily data with large volumes from various learning and teaching activities [6]. The increase in data mining applications on educational data has given rise to the concept of Educational Data Mining (EDM). EDM is

The associate editor coordinating the review of this manuscript and approving it for publication was Laxmisha Rai.

an emerging discipline in which data mining techniques such as statistics and machine learning are used on educational data [7], [8].

EDM is concerned with the development and application of computerized methods to discover patterns in educational datasets that are difficult or impossible to analyze manually due to the large volume of data they contain [1], [9]. From another point of view, EDM is generally applied in the form of developing student models that express students' current knowledge, motivation, metacognition and attitudes [10]. EDM is not limited to this, but it is also effectively used to analyze the data produced by any information system related to learning and education [11]. These data can be related to the interaction of an individual student with the learning system, or they can be very diverse, such as data regarding the collaboration with other students, school administrative data, demographic data, and data regarding students' cognitive and emotional situation [1], [12]. It can be said that research in the field of EDM focuses on discovering useful information for educational institutions to better know and manage their students, as well as to better manage students' learning outcomes and increase their performance [12], [13]. On the other hand, EDM also can be used to design better and smarter learning technology and to better inform learners and educators [14].

Although EDM is a relatively new field of research, it has developed rapidly. EDM has a great transformation potential for factors such as discovering how students learn, predicting learning, and understanding actual learning behavior [14]. As a matter of fact, many EDM studies can be mentioned in the literature as a data mining application on educational data. Examples of these studies can be grouped under the following categories: Predicting students' academic performance [6], [15], [16], [17], [18], [19], [20], learning behaviors [21], students' dropout process, efficiency and quality of teaching such as potential estimation [22], [23], [24], clustering students to extract typical behavioral patterns and estimating students at risk [25], [26], [27], [28],university learning materials and evaluation for curriculum improvement [29] planning and strategy for administrative decision making [30], proposing an EDM framework to support learning [11]. The ultimate goal of these studies is to provide important outputs to improve the quality and delivery of educational systems and propose necessary policies [6], [31].

A. PREVIOUS REVIEWS ON EDM

There are various review articles in the literature that aim to provide a broad perspective on the EDM field at different times. For example, [8] conducted the first study in this field in the early days. In this study, EDM techniques applied in e-learning environments between 1995-2005 were examined. [32] conducted a literature review examining the trends and major changes in EDM research and the reduction in the frequency of relationship mining within the EDM community. [33] reviewed the literature on different stakeholders in education such as students, educators, researchers, institutions, and administrators. The researchers have also provided a list of typical training tasks using EDM techniques. Reference [34] realized a superficial literature review on how data mining can be used for purposes such as student retention and attrition, personal recommendation systems in education, and analyzing lesson management system data. Reference [4] conducted a study to reveal the development in the field of EDM and to organize, analyze and discuss the content of the review based on the results produced by the data mining approach. The content of the study consists of 222 EDM approaches and 18 articles containing EDM tools. Reference [9] carried out a systematic review study of 166 articles published over thirty years (1983-2016) on clustering algorithms and their applicability and usability in the context of EDM. Reference [13] performed a study from a different perspective, examining the most commonly used, accessible, and powerful tools that researchers working in the EDM field can use. [7] conducted a study in which they examined various tasks and applications in the EDM field and categorized them according to their purposes. Reference [35] carried out a review on 72 EDM research articles on the teaching and learning process, considering the educational perspective. Reference [36] conducted a systematic review of 33 articles published in the EDM field between 2007 and the first quarter of 2019. Reference [37] published a new review article in which they updated and enhanced their previous article titled "Data mining in education" from 2013. Reference [38] presented a systematic review of 140 EDM studies related to student performance in classroom learning. Reference [39] conducted a bibliometric analysis of the literature on educational data mining published between 2015 and 2019 (n=194). Reference [40] provided a comprehensive review of machine learning approaches, as well as non-performance factors and characteristics, in three different learning environments (Traditional Learning, Blended Learning, and Online Learning), in a systematic review study of 100 articles. Reference [41] conducted a systematic review of 80 studies from 2016 to 2021 that used EDM methods to predict student performance. Reference [42] provided a detailed perspective on student performance prediction by focusing on approximately 260 studies conducted over the past 20 years, from various perspectives.

B. RATIONALE AND IMPORTANCE OF THE PRESENT STUDY

Many bibliometric analyses, systematic reviews, and survey studies provide a narrow or broad perspective on the EDM field. Although these studies have contributed significantly to the field, there is still a need for studies that provide a broad perspective and reveal the big picture of EDM. Methods such as bibliometric analysis, systematic review, and survey studies can have limitations. The difficulty of studies conducted manually on large data sets can also be included in these limitations [43], [44]. At this point, topic modeling analysis, a machine learning-based approach, stands out. Thanks to topic modeling analysis, automatic information extraction can be performed from large data sets [45]. In this context, topic modeling studies that reveal trends and patterns in a research area and extract hidden patterns have been remarkable in recent years [43], [46], [47], [48], [49], [50]. The lack of a topic modeling study that reveals the big picture of the EDM field and uncovers hidden and semantic patterns in the field makes this study necessary and important. In this context, the current study is important as it is the most comprehensive and first topic modeling-based study in the field. Topic modeling analysis, an innovative approach based on unsupervised machine learning, enables the semi-automatic discovery of hidden semantic patterns from large datasets. The topic modeling approach, which enables computer processing of large data sets, has made it easier to extract hidden semantic patterns in research. In this context, this study, which is the first in the field of EDM, is novel in this respect. In this direction, the current study has examined all studies conducted in the field of EDM from 2007 to the present day and extracted the descriptive characteristics of the field. In addition, research interests and trends of the studies have been explored through Latent Dirichlet Allocation LDA-based topic modeling analysis. It is expected that the outputs of the study will guide researchers in the field.

II. METHODS

This section provides information about the methodology of the study, research questions, data collection process, and data analysis. The study is based on descriptive analysis and LDA-based topic modeling analysis. First, the descriptive characteristics of the studies in the literature were revealed with descriptive analysis, then the hidden patterns in the research were discovered with LDA-based topic modeling, and thus research interests and trends were determined. Bibliometric analysis is used to summarize quantitative statistics such as prominent authors, institutions, journals, subject areas, and research years in publications [51]. Topic modeling analysis is an unsupervised machine learning approach used to automatically extract hidden patterns from large datasets [45]. The topic modeling approach is based on automatically discovering hidden semantic patterns called "topics" from large text datasets [45], [49], [52], [53]. In this study, the LDA algorithm [54], a probabilistic method, was used for topic modeling. LDA-based topic modeling was used because it provides an efficient way to calculate the coherence score used to determine the ideal number of topics [45]. LDA-based topic modeling is effectively used as an innovative approach in many areas, such as natural language processing and literature review of job postings [43], [44], [46], [55]. Figure 1 shows the flow of the developmental stages of this study.

As seen in Figure 1, the research problem was first determined. Following the decision to work on EDM, query criteria were created to access the largest data set. The EDM corpus was obtained with this query. Descriptive and topic modeling analyses were applied separately to this corpus. Descriptive characteristics of the corpus were extracted through descriptive analysis. For the topic modeling analysis, first the title, abstract, and keywords of the articles in the corpus were combined into a single text. Then, by following a number of data preprocessing steps, the data set was made ready for topic modeling analysis. The data set, ready for analysis, was subjected to topic modeling analysis, and topics were discovered. Finally, descriptive analysis results and topic modeling analysis findings are reported and presented.

A. RESEARCH QUESTIONS

The aim of this study is to reveal the big picture of the EDM literature. In this regard, the following research questions were addressed to reveal the details of the studies in the EDM field and to determine research interests and trends:

RQ1: What are the document types and numbers, and distribution of them by year in the field of EDM?

RQ2: Which authors, countries, subject areas, and journals stand out in EDM?

RQ3: What is the distribution of topics of the studies in the field of EDM?

RQ4: How have these topics changed and developed over time?

B. STRATEGIES FOR THE CREATION OF THE CORPUS

The first step towards answering research questions is to create a corpus that includes the EDM literature. In this regard, research studies in the literature have been examined, and it has been seen that the Scopus database is suitable and sufficient for this task. Indeed, Scopus is a widely accepted database used to obtain publications in the highest number related to the field in literature review studies [53], [56], [57]. Scopus is the largest database of abstracts and citations, covering more than 7,000 publishers and over 240 disciplines, including publications on the Web of Science [58], [59]. This feature and its acceptance in the literature have made Scopus the preferred choice. In order to cover the EDM literature, the following primitive query has been created to search for the "educational data mining" group in the abstract, title, and keywords:

TITLE-ABS-KEY ("educational data mining") AND (LIMIT-TO (PUBSTAGE, "final")) AND (EXCLUDE (PUBYEAR, 2023))

This query was executed on 06.03.2023, and all the studies published by the end of 2022 were reached. The query returned a total of 2831 records. The document types of the returned records were examined, and it was decided to include "Conference Paper", "Article", "Book Chapter", "Conference Review", "Review", and "Book" types in the corpus. After this process, a total of 2815 records were obtained. When the distribution of publications by year was examined, it was observed that there were only 23 publications in 2007 and earlier, which is less than 1% of the total number of publications. These records were excluded,

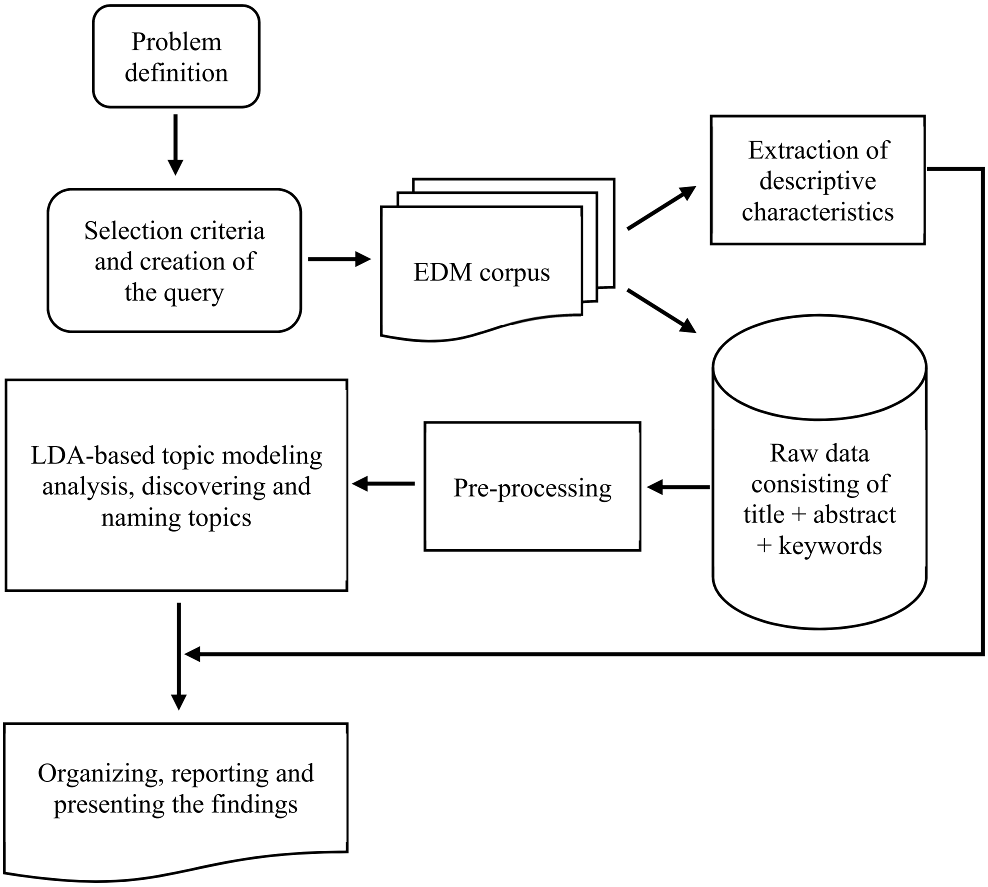

Problem definition
Selection criteria and creation of the query
EDM corpus
Extraction of descriptive characteristics
Raw data consisting of title + abstract + keywords
Pre-processing
LDA-based topic modeling analysis, discovering and naming topics
Organizing, reporting and presenting the findings

FIGURE 1. **Flow chart of the study.**

and the final corpus consisting of the EDM literature from the fifteen-year period between 2008 and 2022 (including 2008 and 2022) was obtained. The final corpus consists of a total of 2792 studies. The bibliometric characteristics of these studies and the title, abstract, and keywords of the corpus were stored in.csv format before the analysis. The final query is as follows:

TITLE-ABS-KEY (“educational data mining”) AND PUBYEAR > 2007 AND (LIMIT-TO (PUBSTAGE, “final”)) AND (LIMIT-TO (DOCTYPE, “cp”) OR LIMIT-TO (DOCTYPE, “ar”) OR LIMIT-TO (DOCTYPE, “ch”) OR LIMIT-TO (DOCTYPE, “cr”) OR LIMIT-TO (DOCTYPE, “re”) OR LIMIT-TO (DOCTYPE, “bk”)) AND (EXCLUDE (PUBYEAR, 2023))

C. PRE-PROCESSING, ADJUSTING THE TOPIC MODELING AND DATA ANALYSIS

The data analysis process of the study consists of two stages. The first stage is the extraction of descriptive characteristics of the EDM literature. In this stage, the obtained data was presented in figures and tables. The second stage is the discovery and naming of topics and trend analysis using LDA-based topic modeling analysis. Topic modeling analysis is basically an unsupervised machine learning technique, also known as a data/text mining approach [60]. Data mining requires some preprocessing steps. The aim of these steps is to get analysis-ready data from raw data. In the preprocessing stage, the combined text consisting of title+ abstract+ keywords of the articles was transformed into plain and clean words. Textual data was converted to lowercase, special characters and punctuation marks were removed, and lemmatization was applied to get the word stems. Then, generic words that do not carry meaning in the text (a, an, the, for, etc.) were added to the stop word list and removed from the text. As a result of these steps, the words in the documents were converted to a word vector according to the “bag of words” logic. All these steps resulted in obtaining cleaned data that is ready for analysis. These operations were carried out using the Python language and data processing libraries.

The data, which was preprocessed and made ready for analysis, was subjected to LDA-based topic modeling analysis. These analyses were also carried out using the Gensim data mining libraries of the Python language [61]. Topic distributions were observed with initial analyses using Gensim’s ldamulticore. The stop word list was checked and additions were made to the list. The words “education”, “data”, “mining”, and “edm” were observed in all topics, and since the research was directly related to this field, it was deemed appropriate to add these words to the stop word list. Then, the final analysis was performed. For each K in the range of

K = [3-25], a model was created in the final analysis. The c_v coherence score was used to determine the ideal number of topics. c_v coherence score is a good solution for determining the ideal number of topics [43], [45]. The topic with the highest c_v coherence score is considered the ideal topic [49], [54]. In Gensim's ldamulticore implementation, the alpha and eta (also known as beta) hyperparameters specify the parameters of the prior Dirichlet distribution. The default values for these two parameters are "symmetric." Various values that these parameters could take (alpha = [symmetric], eta = [symmetric, auto, none]) were tested. The c_v coherence values were obtained for all models. Some important parameters used in the LDA model include "alpha" and "eta". These parameters play an important role in shaping the behavior and output of the model. Alpha is a parameter that controls the generalization of the topic distributions of documents. It determines how the topic distribution in each document will vary. Eta controls the generalization of word distributions representing the content of each topic. This parameter determines which words a topic will contain frequently and which words will be found rarely. The number of K topics determines the number of topics in the model. A c_v value is calculated for each K. In the model, K = [3, 25], a c_v coherence value was calculated for each K. The height of the c_v value is used to determine the ideal number of topics [43], [45]. The results of the experimental trials were examined, and it was determined that the model with K = 12, alpha = "symmetric", eta = "symmetric" provided the highest c_v coherence score (c_v = 0.426). As a result of the analysis, it was decided that the ideal model had 12 topics (K = 12; c_v = 0.426)

After deciding on the ideal number of topics, the topics were visualized using the pyLDAvis library [62], [63]. The visualization was used to name the topics. The lambda value, which shows the importance ranks of the words within the topics, was set to 0.6 as recommended and accepted in the literature [50], [63]. A screenshot of pyLDAvis is given in Figure 2.

Two educational technologists, in addition to the researcher, examined the terms that make up the topics and a consensus was constructed on the final names of the topics. After obtaining the topics and the terms that make them up, a matrix was created showing the publication count for each topic over the years, taking into account the number of publications assigned to each topic. With the help of this matrix, the change of topics over time was traced and trend analysis was carried out.

III. FINDINGS

The findings of the study, in which the fifteen-year-old EDM literature was extensively examined and the hidden patterns of this literature were extracted, are presented with two headings to answer the research questions. The first heading includes the findings related to answers of the first two research questions (RQ1 and RQ2), while the second heading includes the findings related to answers of the third and fourth research questions (RQ3 and RQ4).

TABLE 1. Types of documents that make up the EDM literature, their numbers and percentages.

Document type	n	f
Conference Paper	1566	56.09%
Article	993	35.57%
Book Chapter	107	3.83%
Conference Review	65	2.33%
Review	53	1.90%
Book	8	0.29%

TABLE 2. Top ten authors and their publication numbers in the field of EDM.

Author	n
Baker, R.S.	64
Romero, C.	31
Ventura, S.	23
Nuankaew, P.	21
Leinonen, J.	19
Kotsiantis, S.	18
Hellas, A.	17
Koedinger, K.R.	17
Cechinel, C.	15
Chau, V.T.N.	15

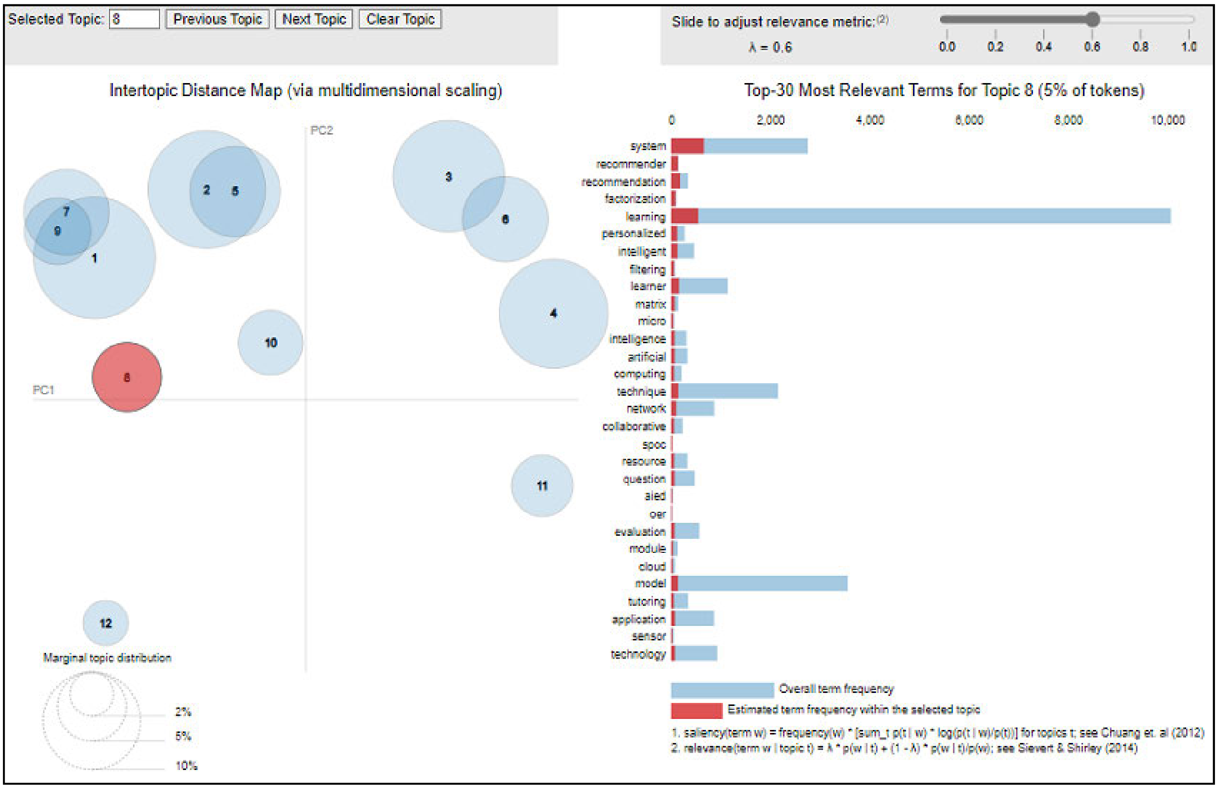
A. FINDINGS ON DESCRIPTIVE CHARACTERISTICS OF THE EDM LITERATURE

In line with the first research question (What are the document types and numbers, and distribution of them by year in the field of EDM?) the document types and numbers and distributions of them by year in the field of EDM literature were determined. While numerical information on document types is given in Table 1, the distribution of the number of documents according to years is given in Figure 3.

As seen in Table 1, more than half of the documents are conference type. The proportion of journal articles (article + review) is 37.9%.

As seen in Figure 3, it can be said that the number of publications in the EDM field has steadily increased over time. This increase continued until 2019 and peaked in that year. Although there was a slight decrease in the number of publications in 2020 compared to the previous year, the number of publications has started to rise again.

In line with the second research question (RQ2: Which are the prominent authors, countries, subject areas and journals in studies in the field of EDM?), the findings regarding prominent authors, countries, subject areas and journals are given in Table 2, Figure 4, figure 5 and Table 3, respectively.



Selected Topic: 8

Slide to adjust relevance metric:(2) λ = 0.6

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

FIGURE 2. A screenshot of pyLDAvis.

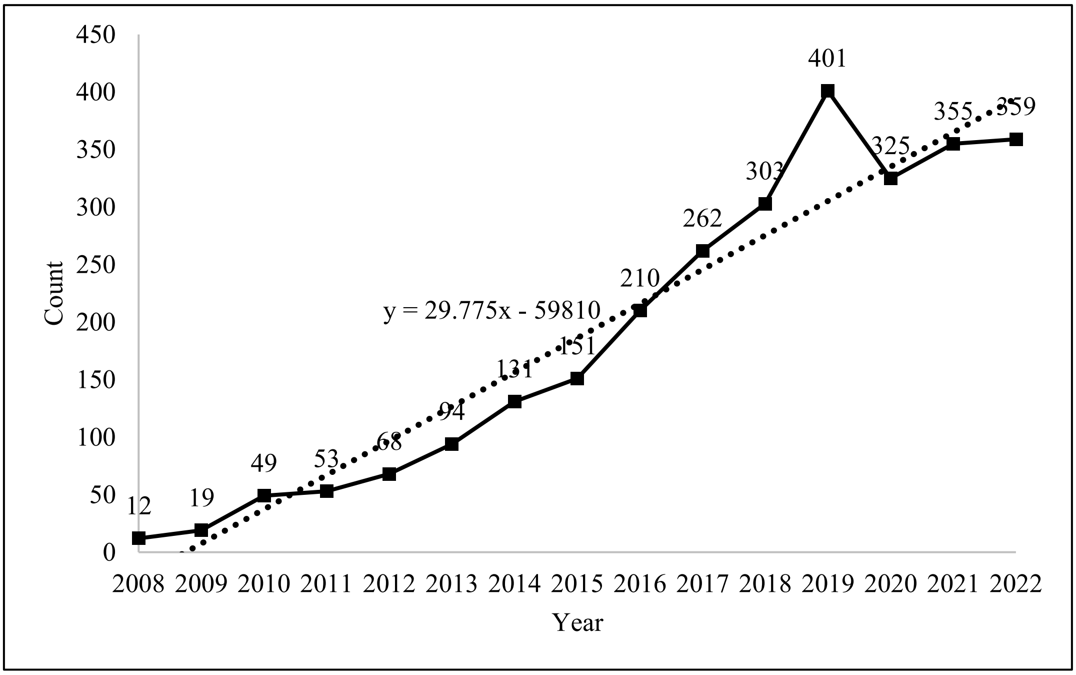


FIGURE 3. Distribution of documents by years and slope line.

As can be seen in Table 2, Baker R.S., Romero C., and Ventura S. are among the most prolific authors in this field (Baker R.S. appears as Baker R.S.J.D in some publications, since they are the same author, the number of publications is summed up and given as one).

As seen in Figure 4, when the origins of the publications are examined, the publications originating from United States, India and China take the lead. In addition, it is seen that countries in different geographies are among the top ten countries.

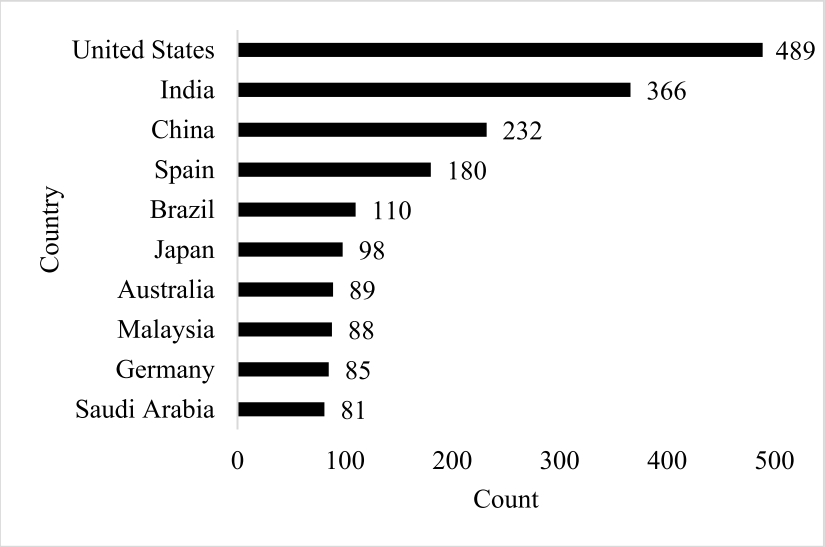


FIGURE 4. Top ten origin countries of publications in the field of EDM and the number of publications.

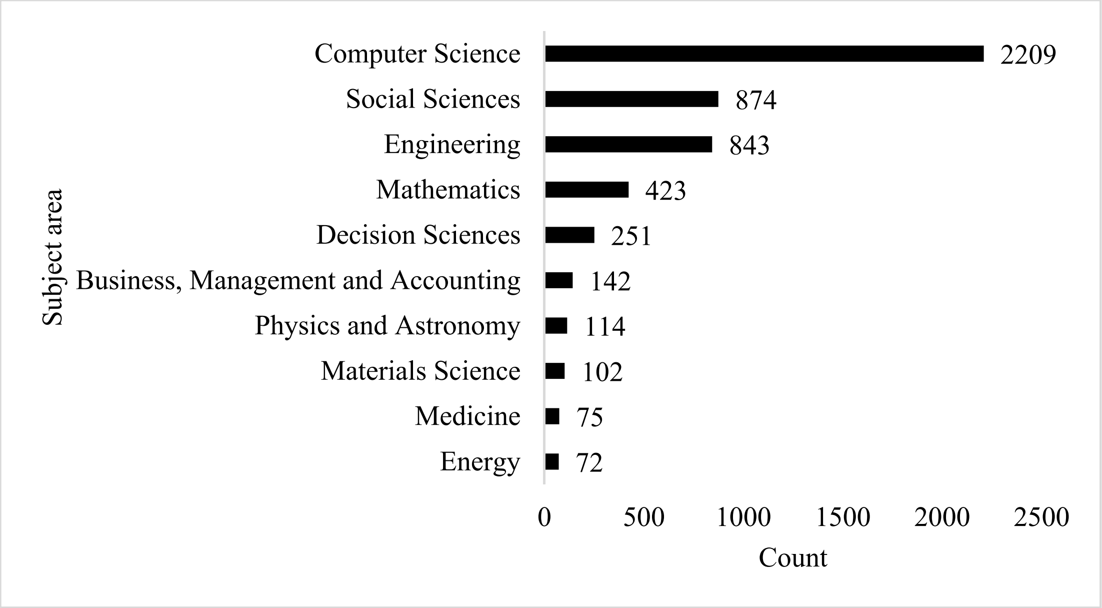


FIGURE 5. Prominent subject areas and number of publications in the field of EDM.

As can be seen in Figure 5, prominent subject areas in publications highlight the interdisciplinary emphasis. As a matter of fact, the top ten subject areas which stand out range from computer science to energy. It should not be misleading that the sum of the subject area publications is more than the total number of publications. This is due to the fact that a post is tagged under more than one subject area.

This subject area classification is an output of the Scopus database. An article is classified into one or more subject areas. The fact that there are different classes (Decision Sciences, Business, Management and Accounting, Physics and Astronomy, Materials Science and Energy, and others) is an indication that EDM-related studies are carried out in different disciplines and fields.

As seen in Table 3, it can be said that the prominent journals in the field are in the fields of computer science and educational technologies.

B. FINDINGS ON TOPIC MODELING ANALYSIS

In this section, the findings related to the emerging topics and their trends in the studies in the field of EDM for answering the third and fourth research questions (RQ3 and RQ4) are given. The results of the analysis revealed that twelve topics emerged in the field of EDM. These topics,

TABLE 3. Top ten journals with the most articles in the field of EDM.

Journal	n
IEEE Access	35
Education and Information Technologies	32
International Journal of Emerging Technologies in Learning	30
International Journal of Advanced Computer Science and Applications	28
Applied Sciences Switzerland	25
IEEE Transactions on Learning Technologies	21
International Journal of Artificial Intelligence in Education	18
Computers and Education	17
Educational Technology and Society	16
Computer Applications in Engineering Education	15

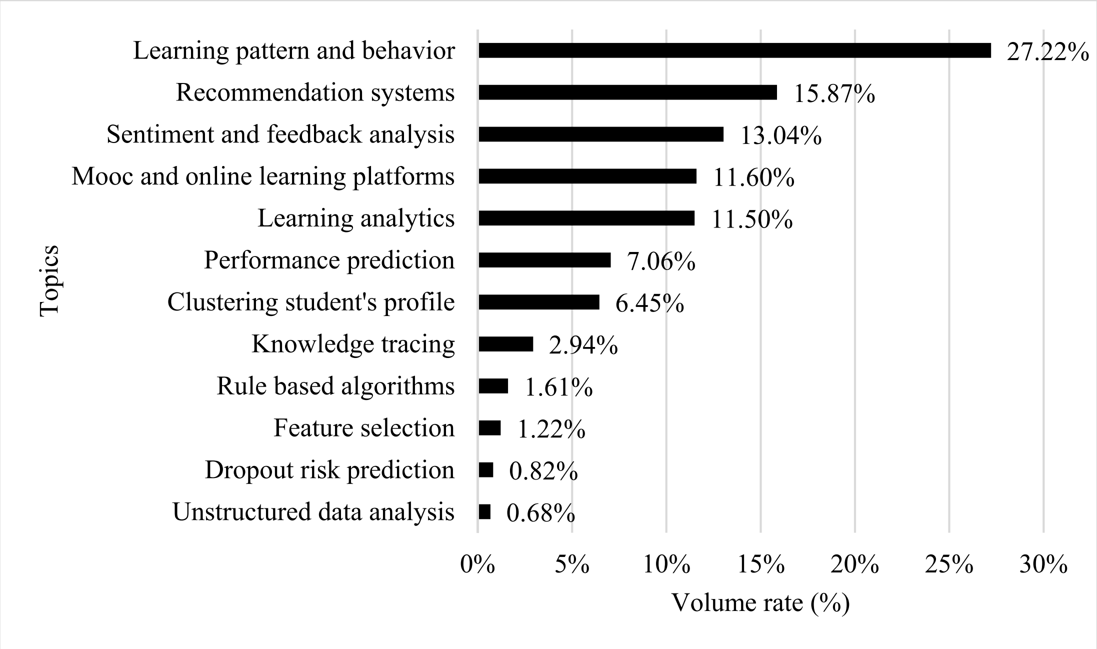


FIGURE 6. The order of the topics according to their volume ratios.

the terms that make up the topics and the volume ratios of the topics are given in Appendix-A. In addition, the number of publications and accelerations of the topics by years are also given in Appendix-B. Firstly, the distribution of topics (for answering RQ3) is listed in Figure 6 in order of volume.

As can be seen from Figure 6, the most voluminous - in other words, the most studied - top three topics in the EDM field are "Learning pattern and behavior", "Recommendation systems" and "Sentiment analysis", respectively. The low-volume topics are identified as "Feature selection", "Dropout risk prediction", and "Unstructured data analysis". The order of the topics by volume ratios and the order of the topics by acceleration are almost equal (can be confirmed from Appendix-A). In fact, when the order of volume is compared with the order of acceleration, it was found that only the topics "Learning analytics (Acc=3.78)" and "Mooc and learning platforms (Acc=2.89)" switched places with each other, and the other ranking remained the same as the order of volume.

To analyze the changes and trends of the topics over time (in response to RQ4), a fifteen-year period has been divided into three-year periods. The percentages of the topics within themselves and compared to other topics over time was obtained by taking into consideration the number of publications in these periods. The basic table where these data were obtained is Table 4, which provides the publication numbers for each period. Accordingly, Table 4 presents the

TABLE 4. Publication numbers of topics in five three-year periods.

Topics \ Periods	2008-2010	2011-2013	2014-2016	2017-2019	2020-2022	Total
Learning pattern and behavior	32	59	130	272	267	760
Recommendation systems	13	39	73	148	170	443
Sentiment and feedback analysis	6	25	57	118	158	364
Mooc and online learning platforms	13	22	72	124	93	324
Learning analytics	4	21	57	113	126	321
Performance prediction	0	11	24	73	89	197
Clustering student's profile	7	18	41	55	59	180
Knowledge tracing	4	14	15	22	27	82
Rule based algorithms	0	3	12	18	12	45
Feature selection	1	0	5	16	12	34
Dropout risk prediction	0	3	2	2	16	23
Unstructured data analysis	0	0	4	5	10	19
Total	80	215	492	966	1039	2792

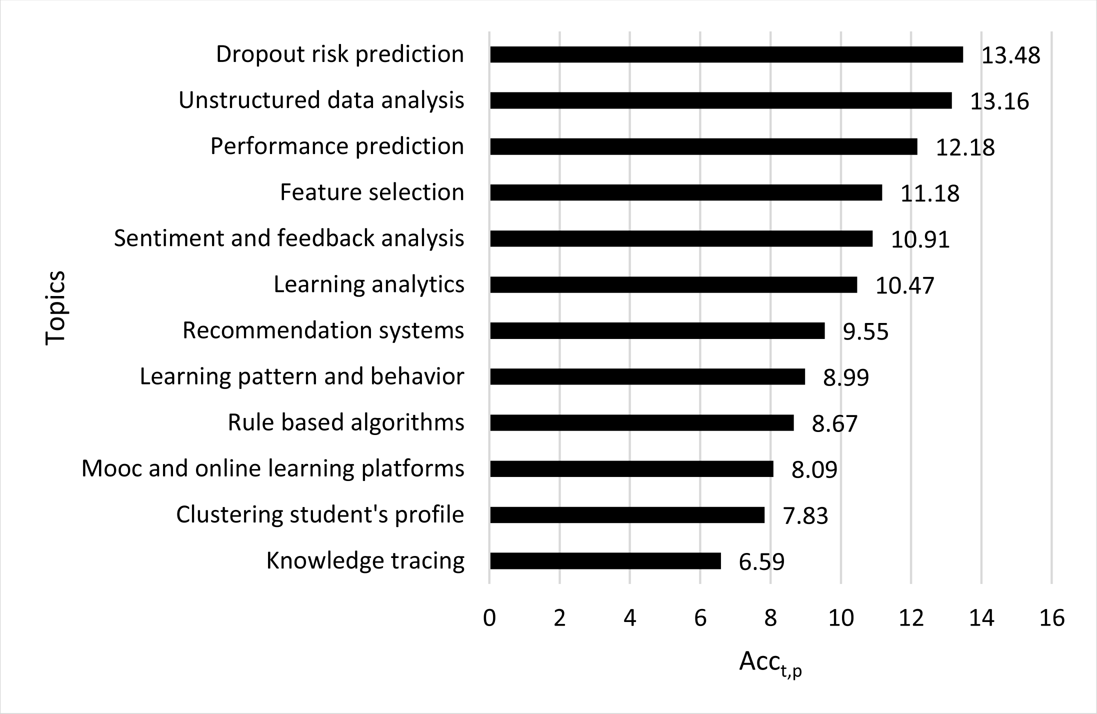


FIGURE 7. The change of the volume ratios of the topics within themselves in periods.

periods and the publication numbers of each topic during these periods.

Using the data in Table 4, the percentage volume of each topic within each periods and the volume percentages of a topic in any period compared to other topics were calculated. For example, in order to calculate the frequency of being studied within itself over time regarding the ‘‘Learning pattern and behavior’’ topic, a row-based reading was performed. Accordingly, the volume ratio of the relevant topic in each period (number of publications in period i/total number

TABLE 5. The volume ratio and acceleration value of each topic in the periods and in comparison to other topics.

Periods	2008-2010	2011-2013	2014-2016	2017-2019	2020-2022	Acc
Topics	$\frac{\textit{Rate in itself}}{\textit{Rate compared to other topic}}$					$\frac{Acc_{t,p}}{Acc_{t,ot,p}}$
Learning pattern and behavior	4.21%	7.76%	17.11%	35.79%	35.13%	8.99
	40.00%	27.44%	26.42%	28.16%	25.70%	-2.79
Recommendation systems	2.93%	8.80%	16.48%	33.41%	38.37%	9.55
	16.25%	18.14%	14.84%	15.32%	16.36%	-0.26
Sentiment and feedback analysis	1.65%	6.87%	15.66%	32.42%	43.41%	10.91
	7.50%	11.63%	11.59%	12.22%	15.21%	1.60
Mooc and online learning platforms	4.01%	6.79%	22.22%	38.27%	28.70%	8.09
	16.25%	10.23%	14.63%	12.84%	8.95%	-1.20
Learning analytics	1.25%	6.54%	17.76%	35.20%	39.25%	10.47
	5.00%	9.77%	11.59%	11.70%	12.13%	1.62
Performance prediction	0.00%	5.58%	12.18%	37.06%	45.18%	12.18
	0.00%	5.12%	4.88%	7.56%	8.57%	1.96
Clustering student's profile	3.89%	10.00%	22.78%	30.56%	32.78%	7.83
	8.75%	8.37%	8.33%	5.69%	5.68%	-0.88
Knowledge tracing	4.88%	17.07%	18.29%	26.83%	32.93%	6.59
	5.00%	6.51%	3.05%	2.28%	2.60%	-0.90
Rule based algorithms	0.00%	6.67%	26.67%	40.00%	26.67%	8.67
	0.00%	1.40%	2.44%	1.86%	1.15%	0.28
Feature selection	2.94%	0.00%	14.71%	47.06%	35.29%	11.18
	1.25%	0.00%	1.02%	1.66%	1.15%	0.15
Dropout risk prediction	0.00%	13.04%	8.70%	8.70%	69.57%	13.48
	0.00%	1.40%	0.41%	0.21%	1.54%	0.19
Unstructured data analysis	0.00%	0.00%	21.05%	26.32%	52.63%	13.16
	0.00%	0.00%	0.81%	0.52%	0.96%	0.24

of publications) was calculated as 4.21%, 7.76%, 17.1%, 35.79% and 35.13%, respectively. Column-based reading was used when calculating the study frequency of this topic compared to other topics in periods. Accordingly, the study frequency of the topics in the first period compared to other topics (i.e., the number of publications on this topic in the first period divided by the total number of publications for that period) was calculated as 40.00%. Similar calculations were performed for all topics, and thus the percentages of each topic's frequency of study over time, both in relation to itself and compared to other topics, were determined. In addition, the acceleration of each topic within each period (Acct,p) and compared to other topics (Acct,ot,p) was also calculated. These data are presented in Table 5.

As can be seen in Table 5, the most frequently studied topic is "Dropuot risk prediction" (Acct,p=13.48), followed by "Unstructured data analysis" (Acct,p=13.16) and "Performance prediction" (Acct,p=12.18), respectively. From another point of view, "Performance prediction" (Acct,ot,p=1.96) was the topic that increased the frequency of study the most compared to other topic over time. This topic is followed by "Learning analytics" (Acct,ot,p=1.62)

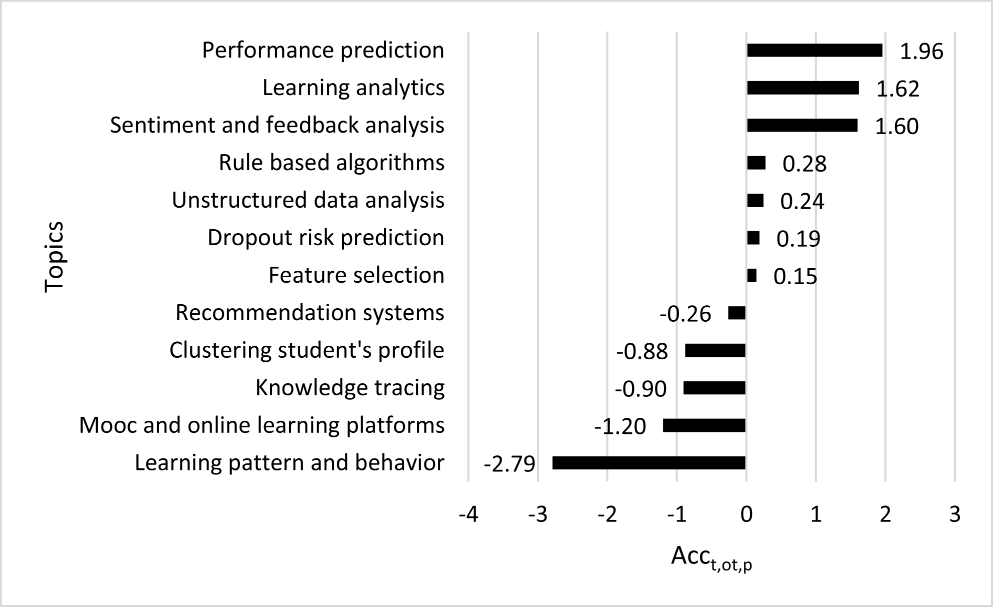


FIGURE 8. Changes in volume ratios of topics compared to other topics over periods.

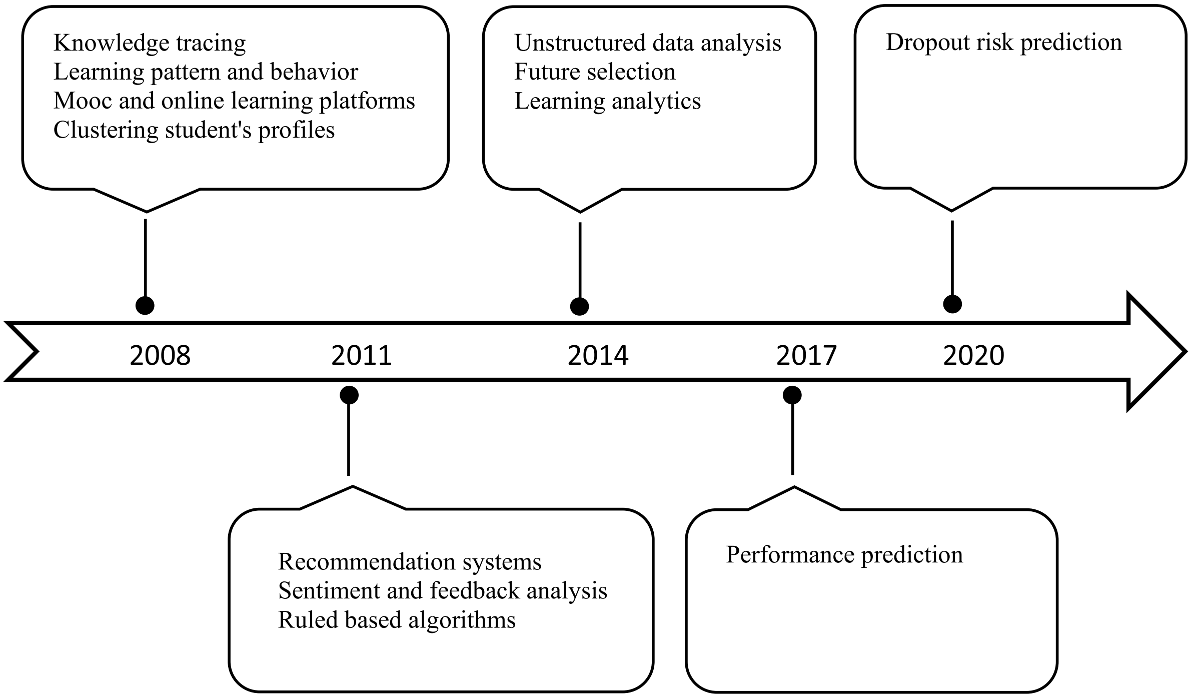


FIGURE 9. Timeline of approximate emergence of topics.

and "Sentiment analysis (Acct,ot,p=1.60)" respectively. Using the data in Table 5, the accelerations of the volume ratios of the topics over time within themselves and relative to other topics are given in figures 7 and 8, respectively.

As seen in Figure 7, "Dropout risk prediction" has been studied more in recent times. In other words, studies on this topic have mostly been carried out in recent periods. This topic is followed by "Unstructured data" and "Performance prediction" topics. The slowest accelerating topic was obtained as "Knowledge tracing".

As seen in Figure 8, while the study frequency of seven topics increased over time compared to other topics, study frequency of five topics decreased over time compared to other topics. While the most prominent topic over time is "Performance prediction", it is followed by "Learning analytics" and "Sentiment analysis" topics. "Learning pattern and behavior" comes first among the topics that are less studied over time compared to other topics, followed by "Mooc and online learning platform" and "Knowledge tracing". Finally, considering the increase in volume ratios of topics over time within themselves, it was also found out when each topic started to come to the fore. In this context, the approximate times when the topics come to the fore have been described and visualized in Figure 9.

As seen in Figure 9, while the topic of "Dropout risk prediction" started to be studied extensively in the 2020s, "Performance prediction" started to gain weight in the 2017's. Thanks to Figure 9, it is possible to see clearly in which years the topics started to become more prominent.

IV. DISCUSSION, LIMITATIONS, AND CONCLUSION

In this section, the results are presented in the light of the findings obtained in the current study and these results are discussed together with the related literature. When the EDM literature was examined, it was seen that conference publications constituted more than half of the corpus. It was observed that the number of documents increased regularly until 2019, and although there was a slight decrease in 2020, it rose again. This situation may be due to interruptions and priority changes in educational researches caused by the Covid-19 pandemic. Indeed, emergency remote education was started with the covid-19 pandemic, studies focused on this area, and interruptions may have occurred in the data processes [64]. Among the most productive authors in this field is Baker, R.S. and United States leads the way in the leading countries. These results are parallel to the literature [39]. On the other hand, it was observed that there are very different fields from "Computer Science" to "Energy" and "Medicine" when the subject areas of the studies were examined. When the subject area categories of the Scopus database are examined, it is seen that the field of "Computer Sciences" takes the lead, still it is possible to say that EDM studies are carried out in many different disciplines. EDM is a field located at the intersection of different disciplines such as computer science, statistics, educational sciences, and psychology. The aim in this field is to use data mining methods to understand student performance by analyzing educational data, improving learning processes, and optimizing educational policies. While computer science provides tools for data analysis, other sciences contribute to the interpretation of educational data and improve the quality of education. In this context, it is natural that EDM studies, which have an interdisciplinary structure, have found application areas in different disciplines. This confirms the interdisciplinary nature of the field [1], [36]. In parallel, the emergence of different journals in the fields of educational technologies and computer sciences, especially "IEEE Access", can be given as an example of the multidisciplinary of the field.

The topic modeling analysis conducted with the studies in the field of EDM gathered these studies under twelve topics. The top three topics, based on volume, are "Learning pattern and behavior," "Recommendation systems," and "Sentiment and feedback analysis." The volume value of these topics also indicates that they are the most studied topics in the field. Numerous studies investigating students' learning patterns, behaviors and strategies in EDM studies draw attention [65], [66]. In addition, the increase in learning resources and the fact that students get lost themselves in these contents [67], [68] have made personalization and suggestion systems important and necessary. In this context, recommendation systems are an important field of study in EDM [69]. Sentiment analysis is one of the commonly used techniques to express human thoughts and is frequently preferred in educational settings. Therefore, sentiment analysis and student feedback analysis systems, which process students' views and opinions through emotion analysis, are among the most studied topics in EDM [70], [71], [72]. In addition to these topics being the most voluminous -most studied topics- in EDM, "Feature selection", "Dropout risk prediction" and "Unstructured data analysis" topics also emerged as unvoluminous topics. Overall, when the volume ranking and acceleration values were compared, it was concluded that the volume ranking and acceleration of the topics are largely the same.

In order to examine the change and development of studies in the field of EDM over time, the fifteen-year time frame was divided into five three-year periods. During these periods, the volume ratios and accelerations of the topics were determined, and the study frequency of the topics within themselves and compared to other topics was determined. When the percentage ratios and accelerations of the volumes of the topics were examined during these periods, the top three topics that have been studied more frequently in recent years were revealed as "Dropout risk prediction", "Unstructured data analysis", and "Performance prediction", respectively. The first two of these topics are low-volume, and the third one is of medium-volume. The fact that the most voluminous topics are relatively present in all periods and that these low and medium voluminous topics have recently started to be studied more may have triggered this situation. Indeed, the years in which these three topics began to gain weight and jump were 2020, 2017, and 2014, respectively. The increase in recent studies aimed at predicting school dropout in both traditional education and Mooc and online environments is remarkable [73], [74], [75]. The results of the study support this. In addition, the increase in different data sources such as text, image, video and the concept of "unstructured data" that has entered our lives with big data [76], has also been used in the field of EDM in recent years [38]. In addition to these, it is not surprising that "Performance prediction" is also among the most studied topics recently. The tremendous increase in learning data has increased the use of EDM techniques for better understanding and organizing the learning process [38], [77], [78], [79].

Finally, the volume ratios of the topics in the periods were compared with the other topics. In this way, the frequency of studying the topics compared to other topics was calculated. In this case, while seven topics stood out more over time among other topics, five topics lagged behind. The top three topics that stood out the most among other topics are ''Performance prediction,'' ''Learning analytics,'' and ''Sentiment and feedback analysis,'' respectively. These topics are the top three topics that gradually increase in weight compared to other topics. The first and third of these are among the most studied and the most voluminous topics in time, respectively. The topic of ''Learning analytics'' ranks sixth among the most studied-on topics over time and fifth in terms of volume. This topic, which started to gain weight in the 2014s, is the second most prominent topic compared to other topics. Learning analytics, defined as measuring, collecting, analyzing, and reporting data about students and contexts to understand and optimize learning environments [80], is used to provide insights into learning processes [81], [82], [83], [84]. In this context, it is not surprising that the topic of ''Learning analytics'' stands out among other topics.

By visualizing the topics with PyLDAvis, the relationship between the topics in the EDM field was seen more clearly. The size of the circles representing the topics indicates the volume and prevalence of the topic. Accordingly, according to the pyLDAvis output, the top three most voluminous topics are represented by the largest circles, and they are the topics ''Learning pattern and behavior'', ''Recommendation systems'', and ''Sentiment and feedback analysis'', respectively. On the other hand, the relationships between the subjects also emerge through the positions of the circles. The distance between circles indicates the similarity or difference between subjects. Accordingly, topics numbered 1-7-9 stand out as close and related topics. These topics were obtained as ''Learning pattern and behavior'', ''Clustering student's profile'', and ''Rule based algorithm''. These three topics are the first group of topics that are related to each other. In addition, topics 2 and 5 (''Recommendation systems'' and ''Learning analytics'') and topics 3 and 6 (''Sentiment and feedback analysis'' and ''Performance prediction'') are close and related topics. Topic number 12 (''Unstructured data analysis'') draws attention as the topic that has the least relationship with all the topics.

This study aims to identify trends in the EDM literature from the past to the present. The study is unique in that it is the first to identify research interests and trends in the EDM literature using an innovative method, topic modeling analysis. However, the study has a number of limitations. The first limitation is that the corpus consists of journal articles only. In future studies, all document types, such as conference proceedings, book chapters, etc., can be included, and topic modeling can be applied to a more comprehensive dataset. Another limitation is the use of the LDA algorithm. The LDA algorithm is an efficient method for topic modeling and is frequently used in such studies. However, in future studies, experimental studies can be conducted with different algorithms, and the results can be presented comparatively. Another limitation is that in topic modeling-based approaches such as LDA, topic naming is done from the authors' point of view and interpretation. On the other hand, it is important to conduct such studies in the future to see how the field has developed. In addition, although this study is the first of its kind, such automated text mining-based research should be encouraged in the future. In this way, the change and development of existing topics and the emergence of new research areas can be observed. Another limitation of the study is a specific situation specific to the field. Since topic modeling is domain-dependent, the emerging topics may be from different research areas, such as tasks or threads. In this case, the topics discovered by taking the context into consideration can be classified at a higher level, and different perspectives can be revealed.

V. IMPLICATIONS

A. FUTURE IMPLICATIONS IN LIGHT OF THE CURRENT SITUATION

The big picture of the EDM field was revealed through the current study. The most voluminous topics in this field and the topics that have been increasingly studied over time both within themselves and in comparison to other topics have been identified. According to the results of the study, the topics with the highest increase in frequency of study over time (the top five topics in terms of growth rate) are low-volume topics such as ''Dropout prediction'', ''Unstructured data analysis'', and ''Feature selection'', as well as ''Performance prediction'' and ''Sentiment and feedback analysis''. In addition, the topics with increasing frequency of study compared to other topics (also the top five topics in terms of growth rate) are ''Performance prediction'', ''Learning analytics'', ''Sentiment and feedback analysis'', ''Rule based algorithms'', and ''Unstructured data analysis'', respectively. Three of the top five topics in both categories (both in itself and in comparison to others) are the same. In addition to high-volume topics, the development of low-volume topics that stand out both within themselves and among other topics should be monitored in the next three to five years. The importance of the current study is evident in terms of understanding the current state and evolution of EDM studies, which is an emerging field. In the light of the current study, similar studies to be conducted in the future will also be important in revealing the evolution of the field. The outputs of current and similar studies are important in terms of guiding both researchers studying in this field and curriculum and policy makers.

B. IMPLICATIONS FOR EDUCATORS AND RESEARCHERS

In the previous section, the current state of research interests and trends in the EDM literature and future perspectives were outlined. This section focuses on the implications for educators and researchers in light of the current results. From the perspective of educators, EDM is known to be used to design better and smarter learning technologies. As a result, learners and educators are better informed. In this context, EDM can be seen as a good tool for educators to make better

TABLE 6. Featured topics in the field of edm, top fifteen terms representing topics and volume ratios.

Topics	First fifteen terms representing the topic	Volume Rate
Learning pattern and behavior	learning, student, system, process, course, analysis, pattern, lm, teacher, behavior, programming, management, information, curriculum, activity	27.22%
Recommendation systems	system, recommender, recommendation, learning, factorization, personalized, intelligent, filtering, learner, matrix, micro, technique, intelligence, artificial, computing	15.87%
Sentiment and feedback analysis	sentiment, teaching, analysis, feedback, big, technique, evaluation, university, higher, student, decision, institution, knowledge, quality, privacy	13.04%
Mooc and online learning platforms	learner, learning, course, moodle, mooc, programming, online, concept, assignment, material, visualization, content, tool, forum, activity	11.60%
Learning analytics	learning, analytics, environment, technology, big, game, online, virtual, tool, learner, analysis, teaching, assessment, student, systematic	11.50%
Performance prediction	performance, student, prediction, machine, algorithm, classifier, model, classification, network, neural, academic, tree, ensemble, decision, technique	7.06%
Clustering student's profile	cluster, pattern, srl, game, online, analysis, learning, hierarchical, interaction, collaborative, gaming, profile, behavior, collaboration, social	6.45%
Knowledge tracing	knowledge, student, tracing, skill, model, cognitive, question, inquiry, tutoring, learning, item, task, intelligent, assessment, response	2.94%
Rule based algorithms	rule, student, performance, algorithm, academic, classification, association, decision, technique, model, admission, prediction, tree, database, clustering	1.61%
Feature selection	feature, selection, classification, algorithm, accuracy, random, dataset, forest, performance, subset, swarm, multilayer, classifier, pso, particle	1.22%
Dropout risk prediction	student, dropout, performance, course, prediction, academic, risk, early, university, model, rate, failure, retention, factor, machine	0.82%
Unstructured data analysis	image, twitter, audio, big, architecture, monitoring, detection, application, resource, blockchain, recognition, bigguery, metadata, iot	0.68%

TABLE 7. Distribution and acceleration of the number of documents pertaining to each topic by years.

	Years																
Topics	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	Total	Acc
Learning pattern and behavior	4	11	17	12	20	27	35	44	51	66	92	114	86	90	91	760	7.66
Recommendation systems	2	1	10	12	16	11	17	26	30	52	41	55	59	60	51	443	4.59
Sentiment and feedback analysis	0	2	4	6	7	12	19	14	24	31	40	47	45	64	49	364	4.38
Mooc and online lerarning platforms	3	2	8	11	5	6	14	24	34	46	35	43	30	28	35	324	2.89
Learning analytics	0	0	4	3	6	12	19	15	23	26	39	48	36	42	48	321	3.78
Performance prediction	0	0	0	2	3	6	8	5	11	7	23	43	28	34	27	197	2.72
Clustering student's profile	2	1	4	5	5	8	9	11	21	18	14	23	22	13	24	180	1.60
Knowlegde tracing	0	2	2	2	6	6	6	2	7	7	5	10	4	12	11	82	0.64
Rule based algorithms	0	0	0	0	0	3	1	7	4	6	6	6	6	3	3	45	0.43
Feature selection	1	0	0	0	0	0	2	2	1	3	6	7	3	4	5	34	0.42
Dropout risk prediction	0	0	0	0	0	3	1	0	1	0	0	2	3	4	9	23	0.37
Unstructered data analysis	0	0	0	0	0	0	0	1	3	0	2	3	3	1	6	19	0.30
Total	12	19	49	53	68	94	131	151	210	262	303	401	325	355	359	2792	

Acc: Acceleration

inferences. Considering that ‘‘Learning patterns and behavior’’, ‘‘Recommendation systems’’, and ‘‘Sentiment and feedback analysis’’ are the most studied topics in the context of the results of the study, it is thought that educators can frequently work on these topics. On the other hand, it can be expected that ‘‘Dropout risk prediction’’, ‘‘Learning analytics’’, and ‘‘Performance prediction’’, which have recently come to the fore, will be the focus of educators’ attention in the near future.

In the context of researchers, the results of the study can be expected to provide important outputs and perspectives. Both the identification of the most studied topics and the topics that have come to the fore in recent years offer important opportunities for researchers in this field in the near future, beyond identifying research interests and trends in this field. In the previous section, predictions for the future were presented in a broad manner. In light of these, it is noteworthy that topics such as ‘‘Dropout prediction’’, ‘‘Unstructured data analysis’’, and ‘‘Feature selection’’, although low in volume, have increased in intensity over time. On the other hand, topics such as ‘‘Performance prediction’’, ‘‘Learning analytics’’, ‘‘Rule based algorithms’’, and ‘‘Unstructured data analysis’’ that stand out compared to other topics may be interesting to follow in the near future.

APPENDIX A

See Table 6.

APPENDIX B

See Table 7.

REFERENCES

[1] C. Romero and S. Ventura, "Data mining in education," *WIREs Data Mining Knowl. Discovery*, vol. 3, no. 1, pp. 12–27, Jan. 2013, doi: 10.1002/widm.1075.

[2] T. Treasure-Jones, C. Sarigianni, R. Maier, P. Santos, and R. Dewey, "Scaffolded contributions, active meetings and scaled engagement: How technology shapes informal learning practices in healthcare SME networks," *Comput. Hum. Behav.*, vol. 95, pp. 1–13, Jun. 2019, doi: 10.1016/j.chb.2018.12.039.

[3] J. Mostow and J. Beck, "Some useful tactics to modify, map and mine data from intelligent tutors," *Natural Lang. Eng.*, vol. 12, no. 2, pp. 195–208, Jun. 2006, doi: 10.1017/S1351324906004153.

[4] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, Mar. 2014, doi: 10.1016/j.eswa.2013.08.042.

[5] S. I. McClean, "Data mining and knowledge discovery," in *Encyclopedia of Physical Science and Technology*. CA, USA: Academic, 2003.

[6] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2, Feb. 2019, Art. no. e01250, doi: 10.1016/j.heliyon.2019.e01250.

[7] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years," *Educ. Inf. Technol.*, vol. 23, no. 1, pp. 537–553, Jan. 2018, doi: 10.1007/s10639-017-9616-z.

[8] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Exp. Syst. Appl.*, vol. 33, no. 1, pp. 135–146, Jul. 2007, doi: 10.1016/j.eswa.2006.04.005.

[9] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017, doi: 10.1109/ACCESS.2017.2654247.

[10] S. K. Mohamad and Z. Tasir, "Educational data mining: A review," *Proc.-Social Behav. Sci.*, vol. 97, pp. 320–324, Nov. 2013, doi: 10.1016/j.sbspro.2013.10.240.

[11] Md. M. Rahman, Y. Watanobe, T. Matsumoto, R. U. Kiran, and K. Nakamura, "Educational data mining to support programming learning using problem-solving data," *IEEE Access*, vol. 10, pp. 26186–26202, 2022, doi: 10.1109/ACCESS.2022.3157288.

[12] A. Abu, "Educational data mining & students' performance prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 5, pp. 212–220, 2016, doi: 10.14569/ijacsa.2016.070531.

[13] S. Slater, S. Joksimovic, V. Kovanovic, R. S. Baker, and D. Gasevic, "Tools for educational data mining: A review," *J. Educ. Behav. Statist.*, vol. 42, no. 1, pp. 85–106, Feb. 2017, doi: 10.3102/1076998616666808.

[14] R. S. Baker, "Educational data mining: An advance for intelligent systems in education," *IEEE Intell. Syst.*, vol. 29, no. 3, pp. 78–82, May 2014, doi: 10.1109/MIS.2014.42.

[15] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Comput. Educ.*, vol. 113, pp. 177–194, Oct. 2017, doi: 10.1016/j.compedu.2017.05.007.

[16] C. Beaulac and J. S. Rosenthal, "Predicting university students' academic success and major using random forests," *Res. Higher Educ.*, vol. 60, no. 7, pp. 1048–1064, Nov. 2019, doi: 10.1007/s11162-019-09546-y.

[17] A. F. ElGamal, "An educational data mining model for predicting student performance in programming course," *Int. J. Comput. Appl.*, vol. 70, no. 17, pp. 22–28, May 2013, doi: 10.5120/12160-8163.

[18] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," *Comput. Educ.*, vol. 61, pp. 133–145, Feb. 2013, doi: 10.1016/j.compedu.2012.08.015.

[19] R. Trakunphutthirak and V. C. S. Lee, "Application of educational data mining approach for student academic performance prediction using progressive temporal data," *J. Educ. Comput. Res.*, vol. 60, no. 3, pp. 742–776, Jun. 2022, doi: 10.1177/07356331211048777.

[20] J. Zimmerman, K. H. Brodersen, H. R. Heinimann, and J. M. Buhmann, "A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance," *J. Educ. Data Mining*, vol. 7, no. 3, pp. 151–176, 2015. [Online]. Available: http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/JEDM070/pdf_19

[21] D. Kim, M. Yoon, I.-H. Jo, and R. M. Branch, "Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in South Korea," *Comput. Educ.*, vol. 127, pp. 233–251, Dec. 2018, doi: 10.1016/j.compedu.2018.08.023.

[22] F. Yang and F. W. B. Li, "Study on student performance estimation, student progress analysis, and student potential prediction based on data mining," *Comput. Educ.*, vol. 123, pp. 97–108, Aug. 2018, doi: 10.1016/j.compedu.2018.04.006.

[23] W. Cambruzzi, S. J. Rigo, and J. L. V. Barbosa, "Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach," *J. Univers. Comput. Sci.*, vol. 21, no. 1, pp. 23–47, 2015.

[24] W. Xing, X. Chen, J. Stein, and M. Marcinkowski, "Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization," *Comput. Hum. Behav.*, vol. 58, pp. 119–129, May 2016, doi: 10.1016/j.chb.2015.12.007.

[25] G. Cobo, D. García-Solórzano, J. A. Morán, E. Santamaría, C. Monzo, and J. Melenchón, "Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums," in *Proc. 2nd Int. Conf. Learn. Analytics Knowl.*, Apr. 2012, pp. 248–251, doi: 10.1145/2330601.2330660.

[26] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Comput. Hum. Behav.*, vol. 104, Mar. 2020, Art. no. 106189, doi: 10.1016/j.chb.2019.106189.

[27] S. Hassan, H. Waheed, N. R. Aljohani, M. Ali, S. Ventura, and F. Herrera, "Virtual learning environment to predict withdrawal by leveraging deep learning," *Int. J. Intell. Syst.*, vol. 34, no. 8, pp. 1935–1952, Aug. 2019, doi: 10.1002/int.22129.

[28] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Comput. Hum. Behav.*, vol. 73, pp. 247–256, Aug. 2017, doi: 10.1016/j.chb.2017.01.047.

[29] Y. H. Jiang, S. S. Javaad, and L. G. Golab, "Data mining of undergraduate course evaluations," *Informat. Educ.*, vol. 15, no. 1, pp. 85–102, Apr. 2016, doi: 10.15388/infedu.2016.05.

[30] V. Caputi and A. Garrido, "Student-oriented planning of e-learning contents for Moodle," *J. Netw. Comput. Appl.*, vol. 53, pp. 115–127, Jul. 2015, doi: 10.1016/j.jnca.2015.04.001.

[31] S. Agarwal, "Data mining in education: Data classification and decision tree approach," *Int. J. e-Educ., e-Bus., e-Manag. e-Learn.*, vol. 2, no. 2, p. 140, 2012, doi: 10.7763/ijeeee.2012.v2.97.

[32] R. S. J. D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Min.*, vol. 1, no. 1, pp. 3–17, 2009.

[33] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern., C*, vol. 40, no. 6, pp. 601–618, Nov. 2010, doi: 10.1109/TSMCC.2010.2053532.

[34] R. A. Huebner, "A survey of educational data-mining research," *Res. High. Educ. J.*, vol. 19, pp. 1–13, Apr. 2013. [Online]. Available: http://0-search.proquest.com.millenium.itesm.mx/docview

[35] M. W. Rodrigues, S. Isotani, and L. E. Zárate, "Educational data mining: A review of evaluation process in the e-learning," *Telematics Informat.*, vol. 35, no. 6, pp. 1701–1717, Sep. 2018, doi: 10.1016/j.tele.2018.04.015.

[36] X. Du, J. Yang, J.-L. Hung, and B. Shelton, "Educational data mining: A systematic review of research and emerging trends," *Inf. Discovery Del.*, vol. 48, no. 4, pp. 225–236, May 2020, doi: 10.1108/IDD-09-2019-0070.

[37] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 3, p. e1355, May 2020, doi: 10.1002/widm.1355.

[38] A. Khan and S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," *Educ. Inf. Technol.*, vol. 26, no. 1, pp. 205–240, Jan. 2021, doi: 10.1007/s10639-020-10230-3.

[39] C. Baek and T. Doleck, "Educational data mining: A bibliometric analysis of an emerging field," *IEEE Access*, vol. 10, pp. 31289–31296, 2022, doi: 10.1109/ACCESS.2022.3160457.

[40] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "Student retention using educational data mining and predictive analytics: A systematic literature review," *IEEE Access*, vol. 10, pp. 72480–72503, 2022, doi: 10.1109/ACCESS.2022.3188767.

[41] W. Xiao, P. Ji, and J. Hu, "A survey on educational data mining methods used for predicting students' performance," *Eng. Rep.*, vol. 4, no. 5, May 2022, Art. no. e12482, doi: 10.1002/eng2.12482.

[42] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Educ. Inf. Technol.*, vol. 28, no. 1, pp. 905–971, 2022, doi: 10.1007/s10639-022-11152-y.

[43] F. Gurcan, N. E. Cagiltay, and K. Cagiltay, "Mapping human-computer interaction research themes and trends from its existence to today: A topic modeling-based review of past 60 years," *Int. J. Hum.-Comput. Interact.*, vol. 37, no. 3, pp. 267–280, Feb. 2021, doi: 10.1080/10447318.2020.1819668.

[44] X.-L. Yang, D. Lo, X. Xia, Z.-Y. Wan, and J.-L. Sun, "What security questions do developers ask? A large-scale study of stack overflow posts," *J. Comput. Sci. Technol.*, vol. 31, no. 5, pp. 910–924, Sep. 2016, doi: 10.1007/s11390-016-1672-0.

[45] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, Nov. 2010.

[46] C. C. Ekin, E. Polat, and S. Hopcan, "Drawing the big picture of games in education: A topic modeling-based review of past 55 years," *Comput. Educ.*, vol. 194, Mar. 2023, Art. no. 104700, doi: 10.1016/j.compedu.2022.104700.

[47] F. Gurcan and O. Ozyurt, "Emerging trends and knowledge domains in e-learning researches: Topic modeling analysis with the article published between 2008–2018," *J. Comput. Educ. Res.*, vol. 8, pp. 738–756, Jan. 2020, doi: 10.18009/jcer.769349.

[48] J. Kang, S. Kim, and S. Roh, "A topic modeling analysis for online news article comments on nurses' workplace bullying," *J. Korean Acad. Nursing*, vol. 49, no. 6, pp. 736–747, 2019, doi: 10.4040/jkan.2019.49.6.736.

[49] O. Ozyurt and A. Ayaz, "Twenty-five years of education and information technologies: Insights from a topic modeling based bibliometric analysis," *Educ. Inf. Technol.*, vol. 27, no. 8, pp. 11025–11054, 2022.

[50] B. Yin and C. H. Yuan, "Detecting latent topics and trends in blended learning using LDA topic modeling," *Educ. Inf. Technol.*, vol. 27, no. 9, pp. 12689–12712, 2022, doi: 10.1007/s10639-022-11118-0.

[51] G. A. Ganjihal and M. P. Gowda, "ACM transactions on information systems (1989–2006): A bibliometric study," *Inf. Stud.*, vol. 14, no. 4, pp. 223–234, 2008. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=llf&AN=502957020&site=ehost-live

[52] A. De Mauro, M. Greco, M. Grimaldi, and P. Ritala, "Human resources for big data professions: A systematic classification of job roles and required skill sets," *Inf. Process. Manag.*, vol. 54, no. 5, pp. 807–817, Sep. 2018, doi: 10.1016/j.ipm.2017.05.004.

[53] H. Özköse, O. Ozyurt, and A. Ayaz, "Management information systems research: A topic modeling based bibliometric analysis," *J. Comput. Inf. Syst.*, vol. 63, no. 5, pp. 1166–1182, 2022, doi: 10.1080/08874417.2022.2132429.

[54] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003, doi: 10.1016/b978-0-12-411519-4.00006-9.

[55] O. Ozyurt, F. Gurcan, G. G. M. Dalveren, and M. Derawi, "Career in cloud computing: Exploratory analysis of in-demand competency areas and skill sets," *Appl. Sci.*, vol. 12, no. 19, p. 9787, Sep. 2022.

[56] N. Kushairi and A. Ahmi, "Flipped classroom in the second decade of the Millenia: A bibliometrics analysis with Lotka's law," *Educ. Inf. Technol.*, vol. 26, no. 4, pp. 4401–4431, Jul. 2021, doi: 10.1007/s10639-021-10457-8.

[57] R. Vijayan, "Teaching and learning during the COVID-19 pandemic: A topic modeling study," *Educ. Sci.*, vol. 11, no. 7, p. 347, Jul. 2021, doi: 10.3390/educsci11070347.

[58] P. Mongeon and A. Paul-Hus, "The journal coverage of web of science and scopus: A comparative analysis," *Scientometrics*, vol. 106, no. 1, pp. 213–228, Jan. 2016, doi: 10.1007/s11192-015-1765-5.

[59] Scopus. (2022). *Content Coverage*. [Online]. Available: https://www.elsevier.com/solutions/scopus/how-scopus-works/content?dgcid=RN_AGCM_Sourced_300005030

[60] C. C. Aggarwal and C. X. Zhai, *Mining Text Data*. London, U.K.: Springer, 2013.

[61] S. Prabhakaran, "Topic modeling with Gensim (Python)," Mach. Learn. Plus, Mar. 2018. Accessed: Jun. 10, 2023. [Online]. Available: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

[62] B. Mabey. (2018). *pyLDAvis Documentation*. [Online]. Available: https://pyldavis.readthedocs.io/_/downloads/en/stable/pdf/

[63] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proc. Workshop Interact. Language Learn., Vis., Interfaces*, 2015, pp. 63–70, doi: 10.3115/v1/w14-3110.

[64] G. Oliveira, J. Grenha Teixeira, A. Torres, and C. Morais, "An exploratory study on the emergency remote education experience of higher education students and teachers during the COVID-19 pandemic," *Brit. J. Educ. Technol.*, vol. 52, no. 4, pp. 1357–1376, Jul. 2021, doi: 10.1111/bjet.13112.

[65] G.-J. Hwang, S.-Y. Wang, and C.-L. Lai, "Effects of a social regulation-based online learning framework on students' learning achievements and behaviors in mathematics," *Comput. Educ.*, vol. 160, Jan. 2021, Art. no. 104031, doi: 10.1016/j.compedu.2020.104031.

[66] F. Zhao, G. J. Hwang, and C. Yin, "A result confirmation-based learning behavior analysis framework for exploring the hidden reasons behind patterns and strategies," *Educ. Technol. Soc.*, vol. 24, no. 1, pp. 138–151, 2021.

[67] C. De Medio, C. Limongelli, F. Sciarrone, and M. Temperini, "MoodleREC: A recommendation system for creating courses using the Moodle e-learning platform," *Comput. Hum. Behav.*, vol. 104, Mar. 2020, Art. no. 106168, doi: 10.1016/j.chb.2019.106168.

[68] S. S. Khanal, P. W. C. Prasad, A. Alsadoon, and A. Maag, "A systematic review: Machine learning based recommendation systems for e-learning," *Educ. Inf. Technol.*, vol. 25, no. 4, pp. 2635–2664, Jul. 2020, doi: 10.1007/s10639-019-10063-9.

[69] G. George and A. M. Lal, "Review of ontology-based recommender systems in e-learning," *Comput. Educ.*, vol. 142, Dec. 2019, Art. no. 103642, doi: 10.1016/j.compedu.2019.103642.

[70] R. K. Jena, "Sentiment mining in a collaborative learning environment: Capitalising on big data," *Behaviour Inf. Technol.*, vol. 38, no. 9, pp. 986–1001, Sep. 2019, doi: 10.1080/0144929X.2019.1625440.

[71] A. Onan, "Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach," *Comput. Appl. Eng. Educ.*, vol. 29, no. 3, pp. 572–589, May 2021, doi: 10.1002/cae.22253.

[72] N. Sharma and V. Jain, "Evaluation and summarization of student feedback using sentiment analysis," in *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020*. Singapore: Springer, 2021, pp. 385–396.

[73] C. Bargmann, L. Thiele, and S. Kauffeld, "Motivation matters: Predicting students' career decidedness and intention to drop out after the first year in higher education," *Higher Educ.*, vol. 83, no. 4, pp. 845–861, Apr. 2022, doi: 10.1007/s10734-021-00707-6.

[74] S. Dass, K. Gary, and J. Cunningham, "Predicting student dropout in self-paced MOOC course using random forest model," *Information*, vol. 12, no. 11, p. 476, Nov. 2021, doi: 10.3390/info12110476.

[75] A. A. Mubarak, H. Cao, and W. Zhang, "Prediction of students' early dropout based on their interaction logs in online learning environment," *Interact. Learn. Environ.*, vol. 30, no. 8, pp. 1414–1433, Jul. 2022, doi: 10.1080/10494820.2020.1727529.

[76] S. Eybers and H. Kahtsr, "In search of insight from unstructured text data: Towards an identification of text mining techniques," in *Proc. Int. Conf. Digit. Sci.* Cham, Switzerland: Springer, 2018, pp. 591–603.

[77] L. M. Abu Zohair, "Prediction of student's performance by modelling small dataset size," *Int. J. Educ. Technol. Higher Educ.*, vol. 16, no. 1, pp. 1–18, Dec. 2019, doi: 10.1186/s41239-019-0160-3.

[78] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," *J. Med. Syst.*, vol. 43, no. 6, pp. 1–15, Jun. 2019, doi: 10.1007/s10916-019-1295-4.

[79] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Comput. Educ.*, vol. 143, Jan. 2020, Art. no. 103676, doi: 10.1016/j.compedu.2019.103676.

[80] D. Clow, "An overview of learning analytics," *Teaching Higher Educ.*, vol. 18, no. 6, pp. 683–695, Aug. 2013, doi: 10.1080/13562517.2013.827653.

[81] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics Informat.*, vol. 37, pp. 13–49, Apr. 2019, doi: 10.1016/j.tele.2019.01.007.

[82] D. Ifenthaler and J. Y.-K. Yau, "Utilising learning analytics to support study success in higher education: A systematic review," *Educ. Technol. Res. Develop.*, vol. 68, no. 4, pp. 1961–1990, Aug. 2020, doi: 10.1007/s11423-020-09788-z.

[83] S. N. Kew and Z. Tasir, "Developing a learning analytics intervention in e-learning to enhance students' learning performance: A case study," *Educ. Inf. Technol.*, vol. 27, no. 5, pp. 7099–7134, Jun. 2022, doi: 10.1007/s10639-022-10904-0.

[84] A. Maag, C. Withana, S. Budhathoki, A. Alsadoon, and T. H. Vo, "Learner-facing learning analytic—Feedback and motivation: A critique," *Learn. Motivat.*, vol. 77, Feb. 2022, Art. no. 101764, doi: 10.1016/j.lmot.2021.101764.

OZCAN OZYURT received the B.Sc. and M.Sc. degrees in computer engineering and the Ph.D. degree in adaptive educational hypermedia in mathematics education from Karadeniz Technical University, Trabzon, Turkey, in 1996, 2000, and 2013, respectively. He is currently a full-time Faculty Member and an Associate Professor with the Software Engineering Department, Faculty of Technology, Karadeniz Technical University. His major research interests include the use of artificial intelligence in education, software engineering, data mining, big data analysis, and semantic topic modeling.

HACER OZYURT received the B.Sc. degree from the Department of Computer and Instructional Technologies, Karadeniz Technical University, Trabzon, Turkey, in 2007, and the Ph.D. degree in adaptive educational hypermedia and computerized adaptive testing in mathematics education from Karadeniz Technical University, in 2013. She is currently a full-time Faculty Member and an Associate Professor with the Software Engineering Department, Faculty of Technology, Karadeniz Technical University. Her major research interests include mobile programming, augmented and virtual reality, and data mining.

DEEPTI MISHRA (Senior Member, IEEE) has been working as an Associate Professor with the Department of Computer Science (IDI), Norwegian University of Science and Technology (NTNU), since 2016. She is currently the Head of the Intelligent Systems and Analytics (ISA) research group and the Educational Technology Laboratory, Department of Computer Science. She has extensive international experience and earlier worked at Monash University Malaysia; Atilim University, Turkey; and various institutions in India. Her research interests include empirical software engineering, artificial intelligence, human–robot interaction, human-computer interaction, and sustainability.