Received January 20, 2022, accepted March 24, 2022, date of publication March 30, 2022, date of current version April 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3163253

Analyzing Trends and Patterns Across the Educational Technology Communities Using Fontana Framework

MANUEL J. GOMEZ, JOSÉ A. RUIPÉREZ-VALIENTE, (Senior Member, IEEE), AND FÉLIX J. GARCÍA CLEMENTE
Faculty of Computer Science, University of Murcia, 30003 Murcia, Spain

Corresponding author: Manuel J. Gomez (manueljesus.gomezm@um.es)

ABSTRACT Nowadays, the use of technology in continuously increasing, making a significant impact in almost every area, including education. New areas have gained much popularity in the last years in educational technology (EdTech), such as Massive Open Online Courses (MOOCs) or computer-supported collaborative learning. In addition, research and interest in this area have also been growing over the years. The quantity of research and scientific publications in EdTech is constantly increasing, and trying to analyze and extract information from a set of research papers is often a very time-consuming task. To make this process easier and solve these limitations, we present `Fontana`, a framework that can quickly perform trend and social network analysis using any corpus of documents and its metadata. Specifically, the framework can: 1) Discover the latest trends given any corpus of documents, using Natural Language Processing (NLP) analysis and keywords (bibliometric approach); 2) Discover the evolution of the trends previously identified over the years; 3) Discover the primary authors and papers, along with hidden relationships between existing communities. To test its functionality, we evaluated the framework using a corpus of papers from the EdTech research field. We also followed an open science methodology making the entire framework available in Open Science Framework (OSF) easy to access and use. The case study successfully proved the capabilities of the framework, revealing some of the most frequent topics in the area, such as "EDM," "learning analytics," or "collaborative learning." We expect our work to help identifying trends and patterns in the EdTech area, using natural language processing and social network analysis to objectively process large amounts of research.

INDEX TERMS EdTech, data mining, bibliometrics, NLP, network analysis, topic modeling.

I. INTRODUCTION

Both the volume and availability of scientific publications are constantly increasing: for example, one biomedical publication is published approximately every two minutes [1]. Analyzing and inferring information from research papers is often a very time-consuming step, specially if we have a large corpus or a small team of researchers performing the analysis. Generally, there are several approaches to do this type of analysis [2], [3]: systematic reviews, scoping reviews, or even meta-reviews of multiple review papers, among many others. The task of reviewing literature in a particular field is usually done through "strategic reading" [4], where researchers consider several publications to identify those that address tasks, methods, resources, and concepts of interests, and then read only a selection of those in detail. This typically raises a question of the number of "false negative" manuscripts (i.e., the number of articles of interest that may have been missed when reviewing a specific task in a given field).

Unfortunately, extracting and exploring methodological information requires a systematic understanding of the literature, and in many cases, is performed within a limited context of publications that are being manually reviewed by a reduced team [5]. Specifically, reviewing methodologies that have been used for a given task is a time-consuming process that requires systematic knowledge and understanding of the literature. An automated methodology could provide an alternative approach for exploring large corpus of documents within a certain field, enabling best and common practices within particular communities, without having to invest a

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

very significant amount of time that might not always be available.

A research area that could alleviate this workload could be bibliometrics. Usually, in this area, analyses about specific research areas or communities are performed. However, we also identify some challenges, such as the effective identification of topics, scalability, replicability of the analyses, or the automated analysis of different research communities given a research field. Typical research in bibliometrics usually analyze title terms, author keywords, the number of citations, percentage of papers depending on some conditions [6]. These analyses could provide interesting results, but they could be insufficient to accomplish our objectives. We could benefit from using other techniques to enhance the typical analyses performed in bibliometrics and provide deeper results when analyzing a corpus of papers. In addition, we could develop an even more robust framework that could be used easily in any research field to compare different research communities.

One research area that have experienced a big growth is educational technology (EdTech). EdTech is frequently a congealed form of the idea that education plus new technology is the primary and best solution to social problems [7]. Technology has always changed what people did, and many people said that the printed press changed education. Moreover, as new technologies have emerged, it became possible to represent information and knowledge in many forms, including pictures, animations or graphics. In addition, new ways of communication have emerged, such as videoconferencing, text messages or social networks [8], and nowadays there are many digital technologies available to support teaching and learning, including interactive whiteboards and tablets [9]. Furthermore, the term technology-enhanced learning is used to describe the application of information and communication technologies to teaching and learning [10]. Despite there are still some limitations that contribute to the still-limited application of technology in education [11] (such as economic ones), research and interest in this area have been growing over the years. This increase is an excellent motivation to analyze the current trends in EdTech and see changes and new emerging patterns. In particular, Educational Data Mining (EDM) is concerned with developing, researching, and applying computerized methods to detect patterns in large collections of educational data – patterns that would otherwise be hard or impossible to analyze due to the enormous volume of data they exist within [12].

In this work, we analyze the EdTech area by using a corpus of papers, trying to characterize different research communities in this field using our developed framework. We go a step further by combining the analysis of full-text manuscripts with their metadata. This is in line with the ‘‘NLP-enhanced Bibliometrics’’ approach [13], allowing us to perform quick analyses combining two different sources of information. On the one hand, this allows us to infer trends based on full-text data and keywords. On the other hand, we can compare these trends based on different communities and sources available in the metadata. In addition, we combine the NLP area with network analysis, using the full-text papers to identify citations between papers. We also go a step further in the network analysis area, analyzing interactions between different communities and not only papers or authors. Furthermore, we combine all these methods in a single framework that can be applied in any context to obtain a quick analysis and overview of any corpus of papers belonging to a research field. More specifically, this paper addresses the following objectives:

1) *Framework Development:* Develop a framework that can quickly perform topical, trend, and network analysis using any corpus of research papers and its metadata from a research field that can have multiple research communities.
    a) *Discover the Main Topics Given a Corpus of Documents:* To achieve this objective, we will implement two approaches: one of them based on NLP-driven topic modeling using the papers’ full text, and another one using the keywords from the metadata.
    b) *Discover the Evolution of Said Trends Over the Years:* Using every year available in the corpus of documents, we will see how the discovered topics have been progressing over time.
    c) *Discover the Main Authors and Papers, Along With Hidden Relationships Between Research Communities:* To do this, we will perform a social network analysis using the metadata collected.
2) *Evaluate the Framework on the EdTech Area:* We will use the framework developed to test its functionality using a corpus of documents containing multiple research communities.

The rest of the paper is organized as follows. Section II reviews background literature on NLP, bibliometrics, network analysis and EdTech. Section III describes the methodology applied to develop each stage of the framework. Section IV present the results of our case study applied in the EdTech area. Then, we finalize the paper with discussion in Section V and conclusions and future work in Section VI.

II. RELATED WORK

A. BIBLIOMETRICS

The term ‘‘statistical bibliography’’ seems to have been first used by E. Wyndham Hulme in 1922 when he delivered two lectures in Bibliography at the University of Cambridge, being later published as a book [14]. Although this term has been used several times in the existing literature, the general feeling is that this term has never been satisfactory, as it is clumsy, not very descriptive, and can be confused with statistics itself or bibliographies on statistics [15]. Moreover, the term ‘‘bibliometrics’’ was first used, so far as can be ascertained, in the Journal of Documentation in 1969 [16]. Since then, numerous definitions have emerged: ‘‘the application of mathematics and statistical methods to books and other media of communication,’’ or ‘‘quantitative analyses of the bibliographic features of a body of literature'' are only two of the existing definitions of this area of bibliometrics [17].

Previous studies have used bibliometrics to analyze trends. For example, authors in [18] aimed to explore the status quo, hot topics, and future prospects in the field of e-waste (Electronic waste). They collected data from the Web of Science Core Collection and used tools such as *CiteSpace V*, *Histcite*, and *VOSviewer* to analyze literature information. They presented several analyses: document types and publication language, annual publications and growth forecast, authors and co-cited authors, countries and institutions, and journals and co-cited journals, among others. We see another example of bibliometrics in [19], where authors collected data about graphene research from 1991 to 2010 from the Science Citation Index database, Conference Proceeding Citation Index database, and Derwent Innovation Index database integrated by Thomson Reuters. Publications, subjects, collaborations, times cited, co-words, cluster analysis of the papers and patents were deeply examined, and *Thomson Data Analyzer (TDA)* and *Aureka* software were employed to analyze the papers as well as patents data for knowledge mapping. Furthermore, authors in [20] reviewed the literature growth and author productivity of Blockchain technology research from 2008 to March 2017. 801 articles were retrieved from the Scopus database and analyzed with a bibliometrics approach using different perspective views. We can also see recent research such as the work of Sood *et al.* [21], where authors present a cocitation analysis of participating nations, authored documents, scientific contributors, journals, and co-occurrence analysis of keywords on 37,445 extracted Scopus indexed scientific literature to understand the development of 3-D printing technology.

In addition, previous studies have also considered the combination of several techniques to address this type of analysis. For example, Buitelaar *et al.* [22] used Saffron, a system that provides insights into a research community or organization by analyzing the main topics and the individuals associated with them (people) through text mining on their writings (documents). For identifying the most important people in a given corpus, Saffron considers various measures of expertise to rank individuals, including the relevance of a term for a person, their experience in a domain, as well as their area coverage (i.e., knowledge of sub-topics in a domain). Furthermore, Meyers *et al.* [23] proposed an open-source high-performing terminology extraction system called Termolator, which utilizes a combination of knowledge-based and statistical components. Termolator identifies potential instances of terminology using a chunking procedure, similar to noun group chunking, but favoring chunks that contain out-of-vocabulary words, nominalizations, technical adjectives, and other specialized word classes. Finally, we found that Yang *et al.* [24] proposed a novel probabilistic topic model that can jointly model authors, papers, cited authors and venues in one single unified model. In our work, we go a step further, creating a framework that combines NLP, bibliometrics and network analysis, performing topic modeling, keyword and network analysis (including authors and citations between papers) using any corpus of documents. As compared to previous work, `Fontana` can provide a more complete framework to incorporate additional useful contextual information that can also compare different research communities. It is, therefore, more applicable to multiple applications related to academic network analysis.

B. NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do valuable things [25]. Elizabeth D. Liddy [26] provides a more detailed definition of the term: ''Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis to achieve human-like language processing for a range of tasks or applications.''

One of NLP's features is to analyze large corpus of text and extract meaningful information from them (e.g., TF-IDF, Bag of Words, topic finding). We found some studies that previously used NLP to discover topics. Authors in [27] tried to provide a solution to sentiment analysis and topic detection in Spanish tweets using different NLP techniques, such as stemmers and lemmatizers, n-grams, word types, negations, valence shifters, or different classification methods. However, they did not find any method standing out since most of them provided similar results. Another conclusion that authors obtained is that tweets are tough to deal with, primarily due to their brevity and lack of context. Similarly, when using research papers, abstracts could provide worse results than using full-text manuscripts. Choi *et al.* [28] examined trends in academic research on personal information privacy, collecting 2,356 documents covering journal articles, reviews, book chapters, conference papers, and working papers published between 1972 and August 2015 from the Scopus database. They used LDA to the abstracts of those extracted documents, discovering topics like technology, algorithms, and social networks. Moreover, in [29], authors used topic modeling on both possibilities mentioned previously: abstracts and full-text data, concluding that using full-text data provides better results, especially in a small corpus of documents. In fact, differences are as significant as 90% high-quality topics for full-text data, compared to 50% high-quality topics for abstract data.

Finally, we also found other studies that have previously applied NLP to full-text manuscripts. An example is the work in [30], which used the text that accompanies citations in scientific articles, along with supervised methods to determine the purpose (i.e., author intention) and the polarity (i.e., author sentiment) of citation. Moreover, that use of NLP for mining scientific papers leads us to the research topic on ''NLP-enhanced Bibliometrics'', which aims to promote interdisciplinary research in bibliometrics, NLP, and computational linguistics in order to enhance the ways bibliometrics can benefit from large-scale text analytics and sense mining of papers [13]. In our work, we use bibliometric elements such as authors and keywords and NLP elements (full-text scientific papers and keywords) to quickly and precisely discover topics and trends. In addition, we also use the full-text data to identify citations between papers.

C. NETWORK ANALYSIS

Network analysis is becoming increasingly popular as a general methodology for understanding complex patterns of interaction. It examines actors who are connected directly or indirectly by one or more different relationships. Any theoretically meaningful unit of analysis may be treated as actors: individuals, groups, organizations, communities, states, or countries [31]. Social network analysis has been used in studies of kinship structure, social mobility, citation and co-citation networks, corporate power, international trade exploitation, collaboration structures, and many other areas [32], [33].

Social network analysis uses data about networks to calculate different measures, such as the number of relations a node has and the extent to which the node is a bridge between other nodes. If we look at the properties of the network as a whole, researchers can look at things like the average path necessary to connect a pair of nodes or the extent to which the network is dominated by one central actor (centralization) [34]. These ideas have also been applied with the purpose of analyzing research publications. For example, authors in [33] collected data about information sciences articles, consulting *CSA Sociological Abstracts Database* (SA), *Medline Advanced* and *PsycINFO*. Based on these data, they built a co-authorship social network to discover collaborative relationships between authors, calculating measures like density, degree centrality, closeness, among others. Authors note that co-authorship is not the only way to describe relationships between scientific authors since citation networks, for instance, could reveal other relationships. However, these are not included in their article. Moreover, in [35], the objective was to analyze the authorship of scientific manuscripts on a certain disease published in scientific journals indexed in the *Medline* database from 1940 to 2009 and to develop a social network analysis applied to the co-authorship of scientific papers. Finally, 13,989 papers produced by 21,350 authors were analyzed, identifying 116 research groups (clusters) made up of 585 authors.

Regarding citation networks, many previous studies have performed citation networks based on a specific area. An example is the work shown in [36], where authors included all papers that were published in the first three editions of the International Conference on Learning Analytics and Knowledge (LAK). In [37], based on OLED organic research, bibliographic information about 16,551 papers was collected, and a citation network was built based on these data. Similarly, other works built citation networks based on other research areas (e.g., gamification [38], sustainability [39]). We go a step further in this network analysis by analyzing interactions between different communities and not only papers or authors.

D. EDUCATIONAL TECHNOLOGY

Educational technology (EdTech) refers to the use of tools, technologies, processes, procedures, resources, and strategies to improve learning experiences in various settings, such as formal learning and informal learning. EdTech approaches evolved from early uses of teaching tools and have rapidly expanded in recent years to include such devices and approaches as mobile technologies, virtual and augmented realities, simulations and immersive environments, collaborative learning, social networking, cloud computing, flipped classrooms, and more [40].

Not only is EdTech a multi-disciplinary area, but it is also multifaceted, having a number of dimensions or areas to take into consideration. One of the things that make educational technology such an exciting profession is the diversity of people, problems, needs, technologies, and solutions that are involved [40]. In [41], authors made a content and authorship analysis of the last 50 years of the British Journal of Educational Technology, finding that the number of articles increased (from 202 articles published between 1970–79 to 712 articles between 2010-2018). They also performed a concept map based on the papers' content, finding that, in the last decade, new topics became more critical, such as learning analytics or collaborative learning.

We found several studies that applied bibliometrics, NLP, and network analysis to papers in the EdTech area. For example, authors in [42] performed a bibliometric analysis in the top-ranked journal on educational technology over the past 40 years. Using the Web of Science database, authors retrieved 3,963 articles published by the journal "Computers & Education" during the period 1978–2018, highlighting the collaboration among authors, institutions, and countries/regions in the research, which became increasingly close. In addition to the bibliometric analysis (including keywords, citations, h-index), authors performed a social network analysis considering authors with more than ten papers published.

Gurcan *et al.* [43] collected metadata from 1,925 peer-reviewed journal articles published between 2000 and 2019. They performed a bibliometric analysis and then used the abstract of each paper to build an LDA model to discover trends in the area. The analysis revealed 16 topics, and among these, the topics "MOOC," "learning assessment," and "e-learning systems" were crucial topics in the field, with a consistently high volume. In our work, we present a case study where we apply our framework's functionality to a corpus of papers in EdTech, as we consider that the increasing interest in this area is an excellent motivation to discover the latest trends and emerging patterns. Unlike previous research analyzing general trends in the area, our work also allows us to discover trends over the years and see how trends have evolved in each community separately.

III. FONTANA FRAMEWORK

In this section, we are going to present the entire structure of `Fontana` [44], the framework that we have developed in order to accomplish our objectives. We divided our framework into five different stages: 1) data acquisition, 2) data pre-processing, 3) final data collection, 4) modeling, 5) analysis.

The entire process is represented in Figure 1. As we can see, the first step is to extract the raw data (PDF and metadata), and then parse the PDF files into TXT files. After that, we link the TXT files and the metadata into a single data structure, making it available to continue with the data cleaning and lemmatization. After the corpus is cleaned and lemmatized, we already have the final data collection ready to apply different modeling algorithms (keyword analysis, topic finding, and network models). Finally, we use the results from these algorithms to analyze and extract exciting information regarding our data collection. To develop the framework, we have used Python, which is an interpreted high-level, general-purpose programming language.

A. DATA ACQUISITION

To use the framework, we need two types of data: 1) files corresponding to the papers' full text (PDF); 2) metadata of each paper (CSV or XLS). On the one hand, to download each paper, we can use any database from any publisher (e.g., Springer Link database, ACM Digital Library). On the other hand, to get each paper's metadata, we can use two different databases, Scopus [45] and Web of Science [46]:

- Scopus uniquely combines a comprehensive, expertly curated abstract and citation database with enriched data and linked scholarly literature across a wide variety of disciplines. Worldwide, Scopus is used by more than 5,000 academic, government and corporate institutions, and is the main data source supporting the Research Intelligence portfolio.
- Web of Science is a rich collection of citation indexes representing the citation connections between scholarly research articles found in the most globally significant journals, books, and proceedings in the sciences, social sciences and art & humanities. It also serves as the standard data set underpinning the journal impact metrics found in the Journal Citation Reports and the institutional performance metrics found in InCites.

Our framework will use the following metadata fields:

- "Author(s)": a list indicating each one of the authors in the paper.
- "Document title": the title of the publication.
- "Year": year of publication.
- "Source title": name of source where the document has been published (conference, journal, book…).
- "Abstract": abstract of the indexed publication.
- "Author keywords": keywords provided by authors before publication.

However, if we obtain a metadata file with additional fields, our framework will filter those fields automatically. In addition, the field "Source title" will be a critical one in our framework since it will allow us to compare different communities (different journals, conferences…).

B. DATA PRE-PROCESSING

Once we have the full texts and metadata, we parse every PDF file into a plain text (TXT) file in the first part of this stage. To make that possible, we found different libraries implemented in Python that can make that work. To choose the best library, we followed the next steps:

1) *Search:* We collected five different libraries (namely *slate*, *pdfMiner*, *pdfPlumber*, *pyPdf*, and *PdfToText*).
2) *Parsing Evaluation:* We tested if the library was able to parse the PDF files.
3) *Manual Text Review:* We manually compared some TXT files with the raw PDF files to check the quality of parsing.

Some libraries could not parse into TXT some specific PDF files, and some other libraries produced empty TXT files after the parsing. We found that the best library was *PdfToText* [47], parsing 100% of the papers with high fidelity. Furthermore, it was the only library that was capable of parsing double-column PDF appropriately. Once we obtain the papers in TXT, the next step is to link each paper's plain text to its metadata. We use Python functionalities to make this step automatically, merging the entire manuscript and the metadata in a single data structure by a common identifier. This is done by analyzing each paper's full text's first sentences and comparing it against the paper title in the metadata.

Once the framework has pre-processed the data, the next step is to clean each paper's full text. To do that, we keep only the paper's main body (including the abstract), removing the title, authors, and references from the full text. Afterward, we perform additional cleaning actions by removing, for example, unnecessary URLs, numbers, or additional space characters. Moreover, to apply NLP techniques afterward, we need to define a set of "stop words"(i.e., words that will not be considered in the text analysis). Some examples of stop words are "et," "table," or "Figure," which are common words that appear in every document but do not provide helpful information to the analysis. From now, we can start treating each paper as a "document", since most of the cleaning process has ended. Once the full text is cleaned, we lemmatize every document using *pywsd* library. The lemmatization is the process of converting a word to its base form, and its implementation in *pywsd* works as follows:

1) It tokenizes the string, dividing it into a set of tokens (words).
2) It uses a Part-Of-Speech (POS) tagger to map each word to a POS tag (adverb, noun, adjective, etcetera).
3) It calls the lemmatizer with the token and the POS tag to get the base form of the word.

C. FINAL DATA COLLECTION

In this stage, we already have the complete data collection, which is cleaned and prepared to apply the modeling

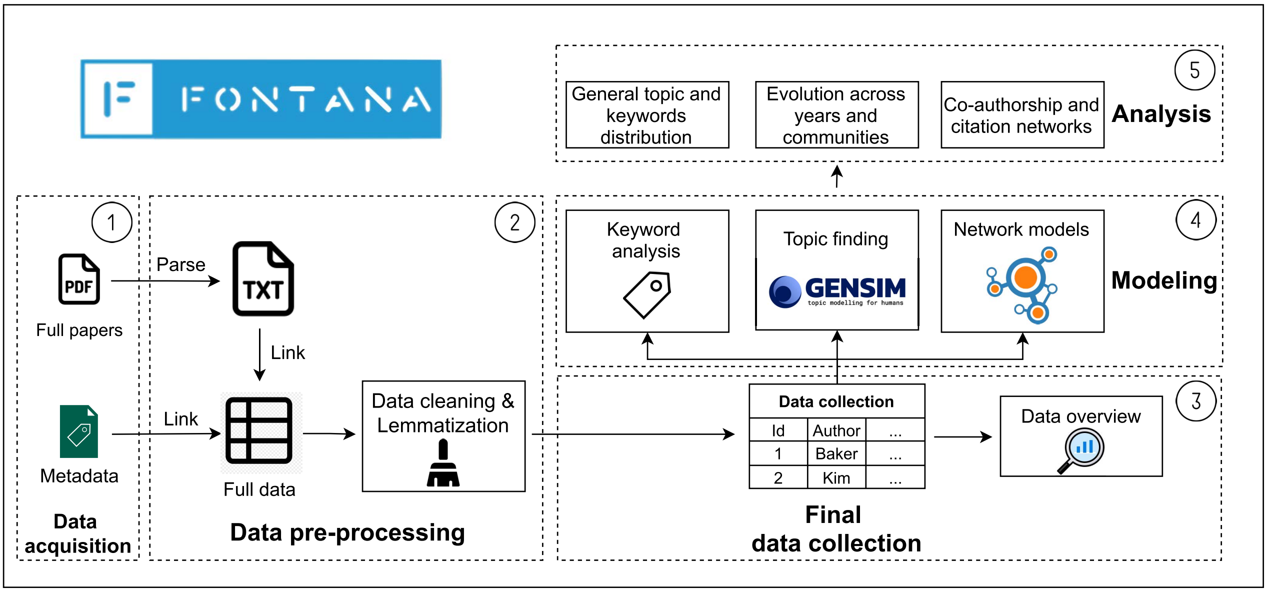


FIGURE 1. **Fontana framework structure.**

algorithms. In order to explore the final data, we can show descriptive statistics, such as the number of documents, the number of different sources (e.g., conferences, journals), the number of words, and the number of unique words. This exploration step is helpful to revise that the final data collection has been processed correctly and includes all the documents expected.

D. MODELING

In this stage, we apply a set of modeling algorithms to our data: 1) Keyword analysis, 2) topic finding, and 3) network models. Although we have selected these three specific modeling algorithms to be applied in our framework, `Fontana` has been designed to be easily expanded with many other algorithms and techniques.

1) KEYWORD ANALYSIS

To discover the main trends in a given corpus based on the papers' keywords, we use all of them collected from the metadata, and then we apply the same cleaning method that we previously applied to the full text. When analyzing papers' keywords, since the authors provide them, some refer to the same term but are slightly different (e.g., technology-enhanced learning, technology-enhanced learning, TEL). Since these similarities can not be found automatically, each user can define similar keywords in the framework and merge them into a single one. This allows us to solve the bias introduced by those keywords and to provide a more reliable analysis.

2) TOPIC FINDING

To discover which are the main topics of a given corpus, we apply Latent Dirichlet Allocation (LDA) topic modeling to the data provided. Previous work [29] has compared two options when performing topic finding on research papers: using the full text or using only the abstract, concluding that using the full text produces a higher number of topics with higher coherence, and therefore better results. For that reason, we have decided to use the full manuscripts in our framework. Specifically, we use *gensim* library and its *ldaMallet* model. Generally in LDA, each document can be described by a distribution of topics, and each topic can be described by a distribution of words; *ldaMallet* uses an optimized Gibbs sampling algorithm for LDA [48].

There are multiple metrics for evaluating the optimal number of topics. Recent studies have shown that the classic predictive likelihood metric (or equivalently, perplexity) and human judgment are often not correlated, and even sometimes slightly anti-correlated [49]. This has led to many studies that have focused upon the development of topic coherence measures. To determine the optimal number of topics, `Fontana` provides two of these coherence measures: C_v and C_{umass} [50], [51]:

- C_v metric is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity.
- C_{umass} is based on document co-occurrence counts, a one-preceding segmentation and a logarithmic conditional probability as confirmation measure.

Based on these two coherence measures, the user can choose the number of topics of the final model calculated. Moreover, to find papers that are strongly related to a particular topic, the framework provides a function that allows users to quickly find those related papers, showing their title, keywords, and identifier. That way, users can find papers related to every topic and have a better idea of which area or areas that topic is referring to.

3) NETWORK MODELS

The next step in the framework is to perform network analysis using both metadata and full-text from papers. Specifically, the framework builds two different networks:

- *Co-Authorship Network:* A co-authorship network in an undirected graph that describes the authors working together within a collection of documents. Each node in the graph represents an author in the collection, and each edge is connected from one author to another that have shared one or more papers. Co-authorship in research articles is considered a reliable proxy of research collaborations, bringing different talents together to give scientific credibility [52].
  In `Fontana`, we use the metadata collected from each paper to build the co-authorship network.
- *Citation Network:* A citation network is a directed graph that describes the citations within a collection of documents. Each node in the graph represents a document in the collection, and each edge is directed from one document toward another that it cites. Since citations of others papers are hand-picked by the authors as being related to their research, the citations can be considered to judge relatedness. Usually, the simplest relation, a direct reference or citation, is likely to occur among related papers which are published apart in time. It does not occur very frequently among papers published in the same year or very close in time [53].
  Since there is no standard format for the citations in bibliographies, and the record linkage of citations can be a time-consuming and complicated process, in `Fontana` we use the references extracted from papers' full-text to locate and represent citations between them. In addition, to identify each paper in the graph, we create an identifier concatenating the first author name with the first word of the paper title and the year of publication (e.g., if the paper is "Student engagement in mobile learning via text message," the first author is "RF Kizilcec," and the year of publication is 2011, the identifier will be "KizilcecStudent2011").

E. ANALYSIS

In this last stage, the framework analyzes the information obtained from the modeling algorithms in order to obtain information of interest. Specifically, we propose three types of analyses: 1) General topic and keywords distribution, 2) evolution across years and communities, and 3) co-authorship and citation networks.

1) GENERAL TOPIC AND KEYWORDS DISTRIBUTION

On the one side, the framework calculates the overall proportion of each keyword as follows:

$$Proportion\_keyword_j = \frac{n\_occurrences_j}{total\_occurrences} * 100 \qquad (1)$$

Then, the proportion of keyword j would be the number of occurrences of j, divided by the total number of occurrences of all keywords ($total\_occurrences$).

On the other side, based on the topics discovered by the topic finding algorithm, the framework calculates the proportion of each topic across the entire corpus. To do that, it evaluates each paper to get its topics associated (note that, in LDA, each document can be assigned to several topics with a certain weight). It calculates the proportion of each topic as follows:

$$Proportion\_topic_j = \frac{\sum_{i=1}^{N} weight\_topic_{ij}}{N} * 100 \qquad (2)$$

Then, the proportion of topic j would be the summation of each weight assigned to the topic j in each document from i to N, divided by the number of documents in the corpus (N). Then, the framework also calculates the proportion of each topic in each source and year individually, using the same equation as shown above, but only using papers of that concrete year and source in each step.

Once the proportions are calculated, the framework provides a visualization showing the proportion of each topic/keyword across the entire corpus. Note that, in the topic proportion visualization, we create a generic identifier in each topic, that will be defined using the three most relevant words of that specific topic. For example, if the most relevant words of a concrete topic are "student,", "problem," and "system," its identifier in the visualization will be "student_problem_system."

2) EVOLUTION ACROSS YEARS AND COMMUNITIES

To discover the evolution of each topic/keyword over the years in each source, the framework also calculates the proportion of each topic and keyword in each source and year individually, using the same equations as shown above, but only using topics/keywords of that concrete year and source in each step. Then, the framework can provide a visualization to discover the evolution across years and communities at a glance.

3) CO-AUTHORSHIP AND CITATION NETWORKS

To build and analyze both networks, we use two different libraries in Python: *Networkx* and *Pyvis*.

Networkx is a Python library/package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [54]. However, *NetworkX* only provides basic and static visualizations since its primary purpose is to perform graph analysis and provide more complex measures. To represent our networks graphically in `Fontana`, we use *Pyvis*, a Python based approach to construct and visualize network graphs in the same space. *Pyvis* provides interactive visualizations, and networks can be customized on a per-node or per edge basis, giving each node different colors, sizes, labels, and other metadata [55].

IV. CASE STUDY RESULTS: EdTech RESEARCH FIELD

A. CONTEXT OF THE CASE STUDY

Since the EdTech area is growing year after year and making a significant impact in educational environments, we consider it ideal for analyzing current trends and discovering changes, new emerging patterns, and hidden relationships between authors and different communities. To select the leading societies within the EdTech area and communities that we are going to compare, we have considered the International Alliance to Advance Learning in the Digital Era (IAALDE). IAALDE represents over 3,000 leading researchers worldwide who have joined forces to advance science, practice, and policy on issues surrounding learning in the context of a digital, technology-driven era [56]. Their objective is to have a more robust global impact by linking together societies with interests focused on the overlap between education, learning, and digital technology. Currently, there are ten different societies involved in this effort, and each individual society has its own history and established mechanisms for supporting the dissemination of their scholarly work. We consider that this alliance embraces research within the EdTech area from several different communities, being very representative of the different trends in this area.

After reviewing the conferences related to each society and checking the availability of their papers, we noted that the papers of the conferences had availability or indexing issues. For example, papers from the International Conference of the Learning Sciences (ICLS), Society for Text & Discourse (ST&D), and International Conference on Educational Technologies (IcEduTech) are not available publicly, and they are not indexed within our databases. Moreover, the Special Interest Group for Building Educational Applications (SIGEDU) society presents workshops instead of conferences. Finally, we also found that papers from International Conference on Computers in Education (ICCE) had indexing issues. Thus, from all the societies part of IAALDE, we were able to select five conferences to analyze their trends, but also to discover the differences between each one of them: Artificial Intelligence in Education (AIED), Educational Data Mining (EDM), Learning Analytics & Knowledge (LAK), Learning at Scale (L@S), and European Conference on Technology Enhanced Learning (EC-TEL). With that purpose in mind, we collected the last five years of each conference and applied our framework to the entire collection of papers.

As we described previously, to use `Fontana`, we need PDF files corresponding to the paper's complete text and the metadata (title, keywords, source, publication year…) of each one of the papers that will be included. On the one hand, to download each paper, we used the different databases corresponding to publishers of each conference (e.g., Springer Link database, ACM Digital Library). We did not collect the complete proceedings since we excluded demo and poster papers from our analysis, and just included either full or short research papers, as including demo or poster papers could introduce some bias due to the papers' reduced size. Those databases contain the papers of each conference for all the editions of the conferences. On the other hand, to get each paper's metadata, we used two different databases, Scopus and Web of Science.

B. ADAPTING THE FRAMEWORK

Although `Fontana` is designed to work with any corpus of papers, each individual case study could require little modifications in order to refine the analysis and solve problems that could emerge. Since we wanted to go a step further in our analysis, we adapted some of the framework's steps manually to provide better and clearer results. Next, we present which steps have been adapted and how we have addressed these modifications:

1) ADDING MISSING METADATA

Using the two databases described above, we collected different CSV and XLS files containing the full metadata needed. We did not find the metadata corresponding to the papers of EDM 2015 edition; thus, we had to include those papers' metadata that we needed for our analysis manually.

2) MERGING KEYWORDS

As we said previously, when the framework analyzes the set of keywords, it can not differentiate between very similar keywords referring to the same term, area or idea. Since we provided the possibility of defining similar keywords to merge all of them into a single one, we used that possibility to define similar keywords in our data:

1) We inspected the 1,000 most frequent keywords to identify similarities between them.
2) We merged every similar keyword into a single one, defining these similarities in a Python dictionary.

For example, we merged the keywords "massive course," "mooc," and "moocs." The result is a single keyword ("Moocs") that aggregates all the occurrences of those similar terms defined.

3) LABELING TOPICS

Since we wanted to assign a representative label to each topic discovered by the model, we followed the next steps to assign this label accurately:

1) We conducted an initial manual topic labeling based on the first five words of each topic.
2) Using the available function that provides strongly related papers to a certain topic, we reviewed ten random papers from each topic to delimit topics more precisely.
3) We assigned the final label to each topic.

In addition, we also modified the plots to show these labels assigned instead of the previous generic identifier created for each topic.

4) VISUALIZING SOCIAL NETWORKS

Although we use interactive and intuitive visualizations in `Fontana` to represent social networks, we wanted to generate a better static visualization for this work. To create these statics visualizations, we have used *Gephi*, an open-source software for graph and network analysis that uses a 3D render engine to display large networks in real-time and allows to speed up the exploration [57]. To integrate *Gephi* with our framework, we use *GephiStreamer*, a Python module that allows to stream graphs directly into *Gephi*. Moreover, to build the citation network, we have collected the metadata from all the available years of the five conferences to obtain a more general and detailed network.

C. DATA COLLECTION OVERVIEW

In this stage, we obtained general information about the corpus, such as the number of documents, the number of different sources (e.g., conferences, journals), the number of words and the number of unique words. Moreover, the framework also shows the most relevant words in the corpus, based on two criteria: 1) Number of occurrences, and 2) Term Frequency-Inverse Document Frequency (TF-IDF), which is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

The final data collection contains a total of 1,334 documents: 50 documents corresponding to 2015, 211 to 2016, 263 to 2017, 270 to 2018, 276 to 2019, and 264 to 2020. Regarding conferences, the corpus has 266 documents corresponding to AIED, 255 to EDM, 233 to L@S, 335 to LAK, and 227 to EC-TEL.

There are a total of 6,695,875 words, being 189,462 of those words unique. Concerning keywords, the entire corpus has 5,580 keywords for 1,334 papers (4.18 average keywords), and 117 papers (8.77%) did not provide any keyword. In Figure 2a we can see the ten most frequent words appearing in our document collection and the number of times that each word appears. Then, in Figure 2b we see the word cloud model applied to our collection. In both models, we can see common words, such as student, use, teacher, data, group or activity. As we expected, the most frequent and essential words are related to learning and technology.

D. GENERAL TOPIC AND KEYWORDS DISTRIBUTION

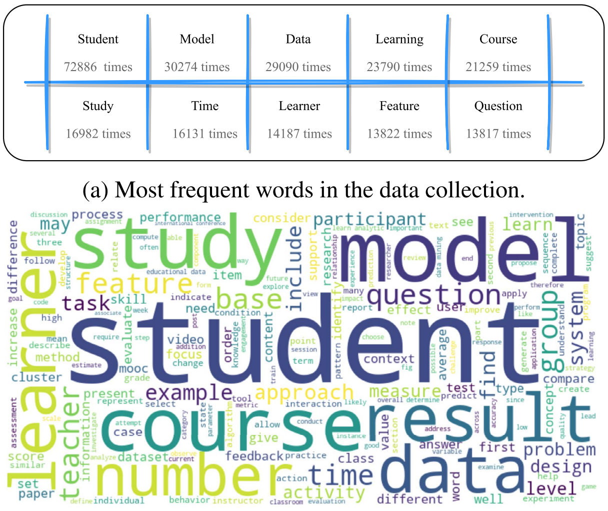
1) USING KEYWORDS

As we know, our framework provides a first approach to discover topics across papers using keywords, which are available in the metadata of each paper.

We can see the distribution of the top-10 keywords across all papers in Figure 3. The most frequent keywords are "learning analytics" (3.4%), "massive open online course" (3.1%), and "intelligent tutoring system" (1.6%). Moreover, the less frequent keywords within the top-10 are "student model" (0.62%) and "educational data mining" (0.77%).

2) USING TOPIC FINDING

After applying the LDA algorithm as described in Section III-D2, we obtained a set of coherence measures to determine which is the optimal number of topics. We determined 18 as the optimal number of topics, with a C_v score of 0.40 and a C_{umass} score of -0.57. A summary of each topic, including its name (assigned following the manual labeling process described in Section IV-B3), description, and the five most important words related, can be found in Table 1.



Student	Model	Data	Learning	Course
72886 times	30274 times	29090 times	23790 times	21259 times
Study	Time	Learner	Feature	Question
16982 times	16131 times	14187 times	13822 times	13817 times

(a) Most frequent words in the data collection.

(b) Most representative words in the data collection.

FIGURE 2. Data collection overview.

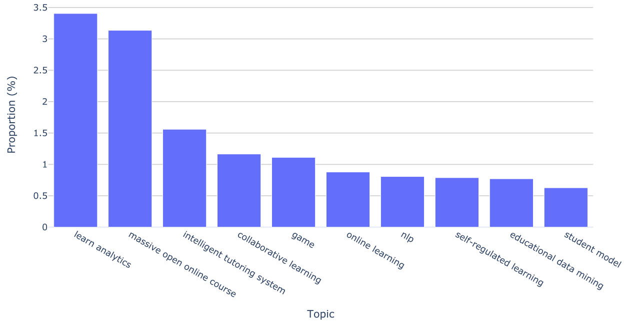


FIGURE 3. Keyword distribution across all papers.

We can see a wide variety of topics, such as games, affective learning, or text analytics. Figure 4 shows the distribution of such topics across the papers. As we can observe, the most frequent topics have been "EDM" (9.9%), "Learner modeling" (8%) and "Learning analytics" (7.5%), and the less frequent topics have been "Multimodal analytics" (3.3%) and "Affective learning" (2.7%). Note that this Figure represents the first global visualization provided by the framework, so it is calculated considering the 1,334 papers of our corpus.

E. EVOLUTION ACROSS YEARS AND COMMUNITIES

The framework also provides us with visualizations showing the evolution over time of topics and keywords discovered. Next, we present the evolution of such topics and keywords in our corpus.

1) USING KEYWORDS

Figure 5 shows the evolution of the most frequent keywords by year and conference. We see some keywords that have

TABLE 1. Summary of each one of the detected topics.

Topic	Description	Main terms
Collaborative learning	Research that promotes the use of groups to enhance learning through working together	Social, network, group, team, collaborative
Knowledge modeling	Papers with the purpose of acquiring new knowledge from existing facts based on certain rules or constraints	Concept, knowledge, domain, expert, rule
Learning at Scale	Research about the technologies, pedagogies, analyses, and theories of learning and teaching that take place with a large number of learners and a high ratio of learners to facilitators.	Learn, online, participant, experience, scale
Clustering	Papers with the task of grouping a set of objects (e.g., students, behaviors) in such way that objects in the same group are similar	Student, cluster, time, pattern, sequence
Experimentation	Includes papers that conduct experimental studies involving learners	Student, score, model, effect, study
Learner modeling	Since every student has individual features such as knowledge, goals or experiences, learner modeling takes advantage of these features and builds a user model or learner model from them.	Student, model, learn, item, problem
Recommendation systems	Papers that use recommender systems applied in education, along with papers aiming to make research accessible (open education)	User, learn, system, recommendation, result
Games	Includes papers that aim to use games and gamification to improve learning	Game, student, learn, level, play
Tutoring systems	Includes papers that aims to provide immediate and customized instruction or feedback to learners, customizing those learning experiences to the unique needs of each individual	Student, condition, tutor, learn, feedback
Multimodal analytics	Focused on analyzing and applying data mining techniques to students' data collected from sensors, such as eye-trackers, motion sensors, or wearables	Video, participant, sensor, device, visual
Learning outcomes	Research focused on the knowledge or skills acquired by learners	Student, grade, exam, high, performance
Computer science education	Focused on computer science competences, such as programming	Student, code, submission, program, feedback
Learning analytics	Learning analytics is defined as the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs [58]	Learn, datum, analytic, research, process
Online learning	Research about online learning (e.g., online courses, massive open online courses)	Video, learn, learner, activity, online
Educational data mining (EDM)	Focused on applying data mining, machine learning and/or statistics techniques to information generated from educational environments	Model, feature, datum, performance, predict
Text analytics	Includes papers that aim to analyze, extract and derive new information from written communication	Question, word, text, topic, score
Visualization dashboards	Papers that develop an understandable and accurate display of information using techniques of data visualization and designing specific dashboards	Student, teacher, design, dashboard, support
Affective learning	Research analyzing learning that is characterized by factors such as motivation, emotions, and other individual psychological aspects of learning	Emotion, affect, student, behavior, confusion

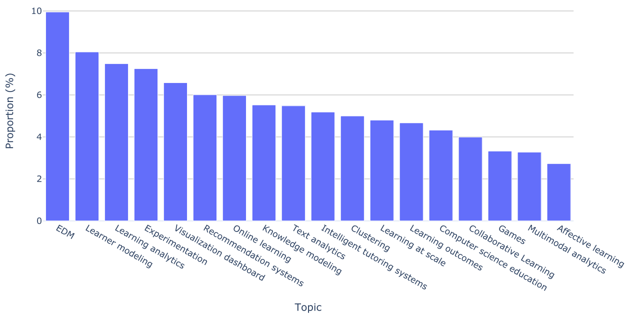


FIGURE 4. Topic's distribution across all papers.

never been as trendy as others, but they keep appearing year after year. This is the case of "student model," which has a stable distribution almost every year in every conference. Moreover, we also see other keywords that have significantly increased their frequency, such as "learn analytics" (increasing from 2.3% in the 2016 edition of EC-TEL to a maximum of 5.88% in 2019). We also see that each keyword does not show a unique trend, as the frequency increase or decrease depends on the different conferences. For example, "massive open online course" increases from 1.8% to 5.7% in EC-TEL, but decreases from 10.3% to 4.3% in L@S.

2) USING TOPIC FINDING

Figure 6 shows the evolution of discovered topics by year and conference. We see that the most frequent topic ("EDM") has a large proportion of appearance in EDM conference (a maximum of 25% in 2020), but it also has a tiny proportion in other conferences, such as EC-TEL (a minimum of 2.4%). Furthermore, we see that the topic "Affective learning"

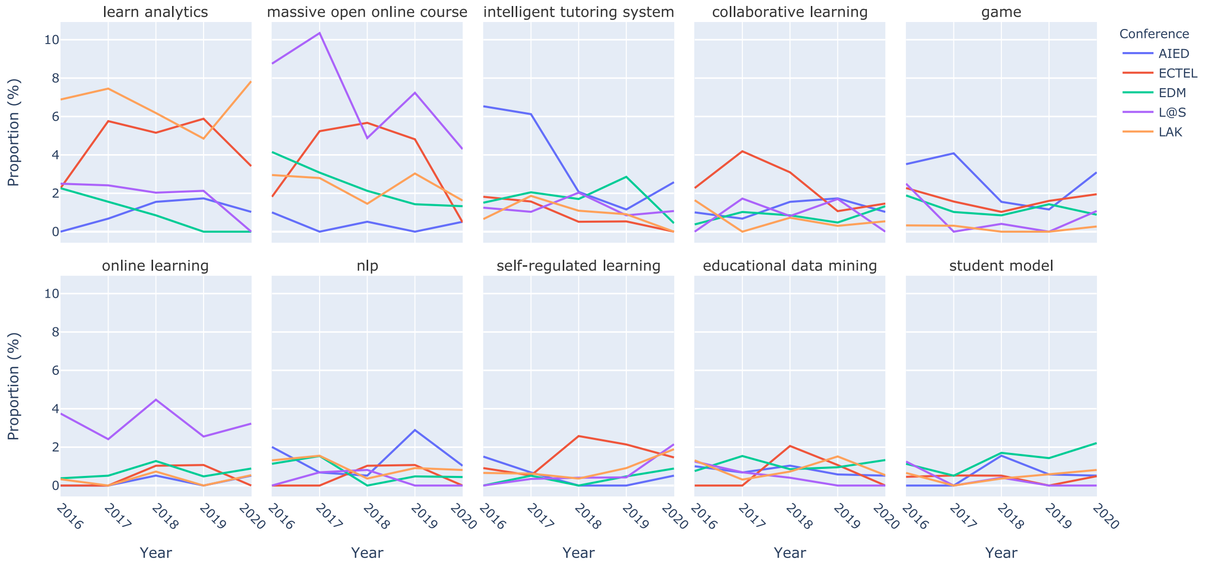


FIGURE 5. Keyword distribution by year and conference.

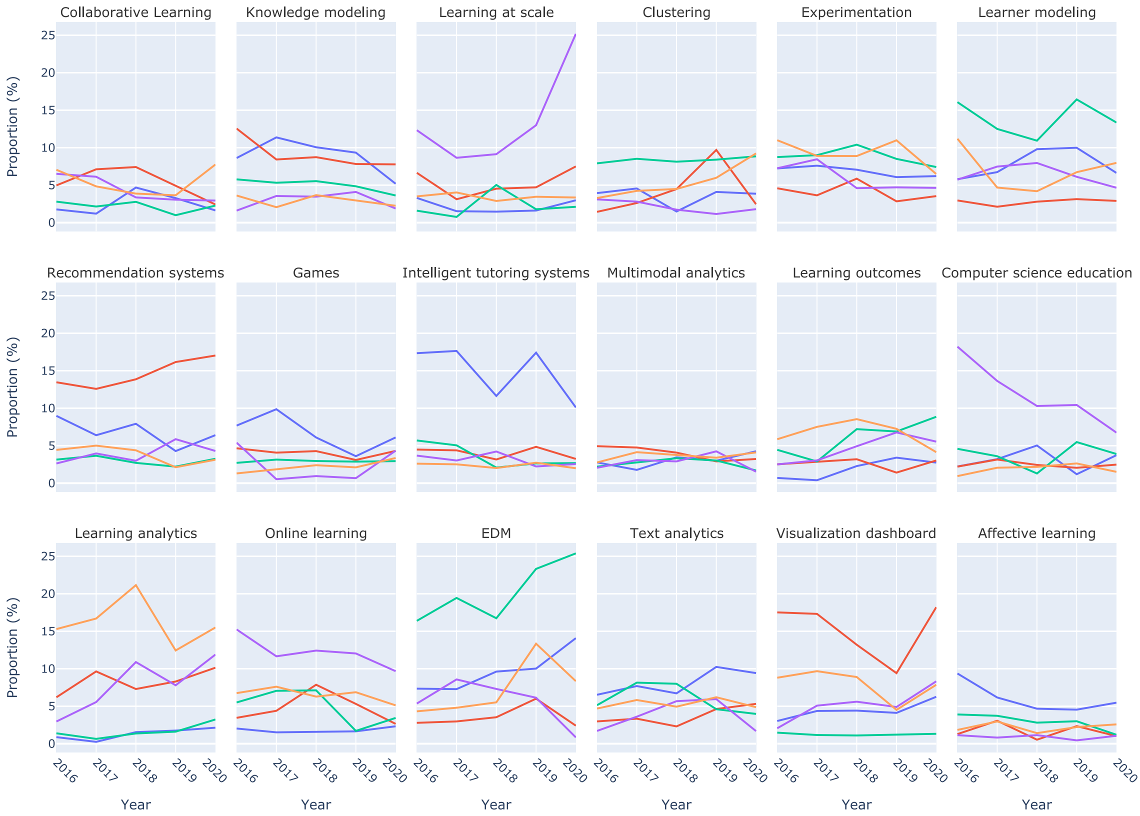


FIGURE 6. Topic's distribution by year and conference.

(the less frequent one) has a small proportion in almost every conference over the years (a maximum of 9.36% in the 2015 edition of AIED).

F. NETWORK ANALYSIS

1) CO-AUTHORSHIP ANALYSIS

In Figure 7, we can see the co-authorship network built with our framework and then streamed into *Gephi*. Each node represents one author in the network, and the color represents the conference where an author has published a larger amount of papers. Note that, in our plot, only the giant component is shown (a giant component is a connected component of a network that contains a significant proportion of the entire nodes in the network).

In total, we have 5,080 author names across papers (an average of four authors per paper), and 2,874 of those

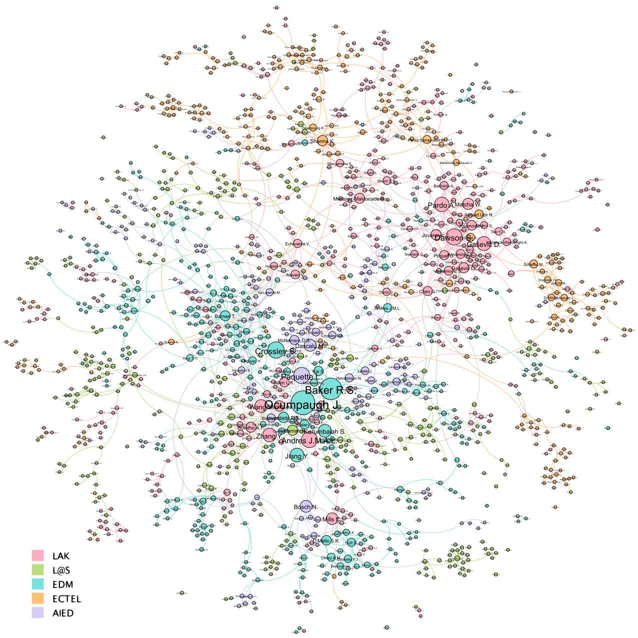


FIGURE 7. Co-authorship network.

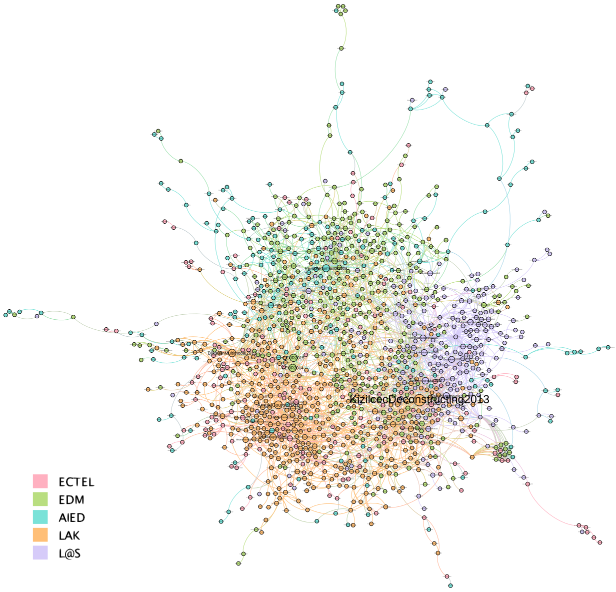


FIGURE 8. Citation network.

author names are unique. Furthermore, the top five central authors are ‘‘Ocumpaugh J.,’’ ‘‘Baker R.S.,’’ ‘‘Joyner D.A.,’’ ‘‘Aleven V.,’’ and ‘‘Gašević D.’’ We also see that, usually, authors that have published a large number of papers in the same conference appear together. We can also see other measures, for example, the percentage of published papers by each author. For example, ‘‘Baker R.S.’’ appears in 3.3% of the papers, while ‘‘Ocumpaugh J.’’ appears in 1% of the papers.

2) CITATION ANALYSIS

In Figure 8, we can see the citation network built with our framework and then streamed into *Gephi*. Each node represents a paper in the network, and the color represents the conference where papers were published. Note that, in our plot, only the giant component is shown.

The framework has found 3,048 references between papers. The top three central papers are ‘‘Students, systems, and interactions: synthesizing the first four years of learning@ scale and charting the future’’ [59], ‘‘Multimodal learning analytics’’ [60], and ‘‘Deconstructing disengagement: analyzing learner subpopulations in massive open online courses’’ [61].

If we look at the results aggregated by conference, we see that the conference that has been cited the most is LAK (1,101 citations), followed by EDM (675 citations) and L@S (549 citations). Moreover, the framework also provides other interesting measures, such as the citations that an average paper from a particular conference has from each one of the conferences. Our results show that the average paper from LAK has 0.98 citations from LAK papers (which means that LAK papers are commonly cited between them), and the same thing happens with L@S. Meanwhile, other conferences show opposite results: the average paper from EC-TEL has 0.09 citations from EC-TEL papers, meaning that EC-TEL papers usually do not cite other papers from the same conference.

G. INTERPRETING RESULTS WITHIN EdTech

As we have seen, our case study has found the main trends in the EdTech area during the last five years, basing our analysis on two primary sources: the full manuscripts and the metadata (authors and keywords). We note that most of the topics revealed by both methods are common (e.g., games, learning analytics, collaborative learning, text analytics), which reaffirms our results' validity. However, some topics' distribution and evolution are different when comparing both methods. For example, in our analysis using the full manuscripts, the topic ‘‘EDM’’ represents 9.9% of the papers, and when using keywords, it only represents 0.77%. The key difference between these two approaches is that the authors carefully select keywords to fit within the current research lines and communities. In contrast, the LDA model is finding hidden topics based on the full texts of the articles, which may reflect more realistically the topic of the paper.

However, not all topics show different distributions and evolution when comparing our two approaches. This is the case of the topic ‘‘Games’’, which represents 3.3% using full manuscripts, and 1.1% when using keywords. In addition, if we look into the evolution of this topic in both approaches, we see that the plot is almost identical. This indicates that authors from this area might be more likely to provide more related keywords in their papers.

As we mentioned before in Section II, some works have previously analyzed topics in this area. For example,

[62] analyzed trends using paper abstracts to conduct LDA topic finding. They found some topics that also appear in our study, like "Collaborative learning" or "Games." However, this study also found some topics that are not so common at present, such as "Blended learning." [63] also analyzed trends in e-learning from 2000 to 2008, conducting text mining and grouping documents based on abstract similarities, and then agglomerating clusters in a hierarchical tree structure. This study revealed some more specific trends such as "Architecture and standards," "Simulations," or "E-learning applications in medical education and training." In this research, we also see some trends that are aligned with our results. For example, the trend "community and interactions" can be associated with our topic "Collaborative learning." Another common result is that new trends like "Knowledge inference" are emerging in this area, which is very positive, as the use of new algorithms and techniques to infer more complex information from the available data is one of the challenges to overcome nowadays.

If we look into the co-authorship network, we see that authors from different communities tend to be mixed, as we do not see any isolated community. Concerning central authors (showing bigger nodes in the network), we note that most are from the LAK and EDM conferences. Then, looking at the citation graph, we see that most of the bigger nodes (most central papers) are, again, from EDM and LAK conferences, but also L@S. We see a little more isolation between communities since we see three primary clusters. The first one is located on the right side of the graph and contains papers from L@S. Then, the other two contain papers from EDM (upper side of the graph) and LAK (left part of the graph). The other two communities (AIED and EC-TEL) do not tend to be as isolated as the rest since their papers are dispersed over the whole graph. Specially, we see a more significant amount of citations between AIED and EDM papers, and then between EC-TEL and LAK papers. Thus, the graph allows us to see a stronger relationship between specific communities (such as AIED and EDM) rather than the rest. With this, we close our case study after confirming that Fontana framework accomplished the established goals within the EdTech research field.

V. DISCUSSION

In Section II we identified some previous studies that also tried to combine several techniques to analyze a research field. For example, [43] performed a bibliometric analysis using a corpus from the e-learning field, and then used the abstract of each paper to build an LDA model to discover existing trends. Moreover, [42] aimed to analyze the research status and trends of the educational technology field, conducting a bibliometric analysis on research topics, author profiles, and collaboration networks. Finally, we also found that [24] proposed a novel probabilistic topic model that jointly models authors, documents, cited authors, and venues simultaneously in one integrated framework, as compared to previous work, which embedded fewer components. In this research, we have gone beyond literature, combining the use of several techniques in order to enhance the classic bibliometrics approach. Combining both full texts and metadata, we can perform quick analyses combining two different sources of information. On the one side, we can perform effective trend identification based on full-text data and keywords. On the other hand, using authors from metadata and citations from full-text data allows us to conduct network analysis and reveal the most central papers and authors given a collection of research papers. Furthermore, we also compare different communities in both trend identification and network analysis, and not only authors or papers. Finally, we have integrated all these methods into a unified and modular framework that can be applied in any research field with a corpus of papers and their metadata.

Another essential feature of our framework is that it is very easy to extend with new functionalities (modeling algorithms), since it has been designed in a modular way. For example, given that we have calculated the citations between different papers (and we also have the authors of each paper), an interesting extension to our framework would be trying to predict the citations of a new paper (i.e., trying to predict which authors would cite a given paper based on which papers they cited previously). The same thing could be done with the number of citations, predicting how many citations a paper would receive based on similar papers. Another example could be to perform social network analysis in each year individually and see the evolution over time of central authors and papers. All these new features could be implemented in our framework very easily, extending it and making it even more complete.

A crucial open challenge at present is replication and transferring the research to practice. It is crucial to provide a detailed description of the procedure followed to conduct the study. While the community is currently demanding more standardized open science practices, this problem is currently still present. This problem of missing information is a familiar issue in multiple research fields (nearly every field is affected), leading to other problems such as low reproducibility. In fact, the terms "reproducibility crisis" and "replication crisis" have gained significant popularity over the last decade [64]. To fix this issue, the community is demanding having more pre-registered studies, open data, open analyses, and open access publications [65], and this can be systematized by the guidelines of the publishers, governments, and research communities [66]. In this research, we have made our framework entirely accessible via an OSF repository. In addition, we have also made the case study presented in this research entirely reproducible [44]. With that purpose, we uploaded the entire database of plain texts after being parsed, the metadata corresponding to each year, and the entire framework code (including scripts and notebooks). The scripts allow to reproduce our work step by step and obtain the same results as we did. Other researchers can easily use this framework to fully re-apply this methodology in other research fields.

This work also has some limitations due to the different analyses that we perform and the several decisions that we have made during the development of `Fontana` and the case study. First, we are limited by authors' keywords, expecting they cover all the possible topics addressed in each paper. However, usually, it is not like that. In addition, the use of these keywords represents another limitation since there might be different keywords referring to the same area or idea that we have not considered. Another limitation is related to the fact that some authors and organizations have different ways to present their names, and this can introduce some bias in the network analysis. Furthermore, another significant limitation is the PDF parsing, since this process could introduce rare characters in the file, also introducing some bias in our analysis. When selecting the number of topics (also in our case study), although we followed a clear methodology in the LDA algorithm, we are also limited by that number since there might be some hidden topics that can not be discovered due to the number of topics selected (a bigger number could reveal more topics). Furthermore, this approach to discover trends and topics can not provide as much in-depth information as other types of qualitative and manual reviews can. However, it offers helpful information that can be enough for multiple purposes, specially within the bibliometrics field. Finally, some methodological limitations can be solved manually as we did in Section IV-B, such as labeling the topics to get a clearer view of what specific research field that topic is describing.

As part of our future work, we will be expanding this analysis to the conferences' entire trajectory, including demo and poster papers. Moreover, since it can be easily done, new modeling algorithms and analyses could be added to the framework to obtain even more interesting and meaningful information. In addition, we would like to improve the framework and create an independent application capable of collecting metadata by itself, so only PDF files would be needed for the analysis.

VI. CONCLUSION

This work aimed to analyze the EdTech area using a corpus of papers, trying to characterize the different communities within the area. Since it has experienced a significant growth over the last years, we consider it an ideal area to analyze its current trends and see changes over time and new emerging patterns. To perform this analysis, we developed a framework capable of performing trend and network analysis using any corpus of documents and their metadata. Furthermore, `Fontana`, our framework built-in Python, was created with three sub-objectives: 1) to discover the latest trends given a corpus of papers; 2) to discover the evolution of such trends over the years; 3) to discover the primary authors and papers, along with hidden relationships between communities. We presented a first approach using the papers' full text and LDA topic modeling, and then we presented a second approach using keywords provided by the authors. Then, we discovered the evolution of said topics over the years using a set of visualizations that allowed us to represent the proportion of each topic in each conference edition. Then, we introduced how the framework can perform social network analysis to show relationships between authors (co-authorship network) and between papers (citation network). Using these same networks, we could also find the main authors and papers of the collection.

Thus, this work provides significant contributions to the literature, including a framework that is scalable, quick, and can be easily be applied in any research field to perform trend and social network analysis. The case study in the EdTech area successfully proved the framework's capabilities, revealing interesting trends and relationships between different research communities in this field. In addition, since we developed the framework in a modular way, `Fontana` can be easily expanded with new analyses and methods to provide new information. We also followed an open data methodology [44] in order to make our framework easy to access and use. That way, other researchers could benefit from our work and make similar analyses in many other contexts.

REFERENCES

[1] J. P. DeShazo, D. L. LaVallie, and F. M. Wolf, "Publication trends in the medical informatics literature: 20 years of 'medical informatics' in MeSH," *BMC Med. Informat. Decis. Making*, vol. 9, no. 1, pp. 1–13, Dec. 2009.

[2] H. Noble and J. Smith, "Reviewing the literature: Choosing a review design," *Evidence-Based Nursing*, vol. 21, no. 2, pp. 39–41, 2018. [Online]. Available: https://ebn.bmj.com/content/21/2/39, doi: 10.1136/eb-2018-102895.

[3] M. Rickinson and H. May, *A Comparative Study of Methodological Approaches to Reviewing Literature*. Higher Education Academy, 2009.

[4] address = "United Kingdom",O. Díaz, J. P. Contell, and J. R. Venable, "Strategic reading in design science: Let root-cause analysis guide your readings," in *Proc. Int. Conf. Design Sci. Res. Inf. Syst. Technol.* Cham, Switzerland: Springer, 2017, pp. 231–246.

[5] A. Kovačević, Z. Konjović, B. Milosavljević, and G. Nenadic, "Mining methodologies from NLP publications: A case study in automatic terminology recognition," *Comput. Speech Lang.*, vol. 26, no. 2, pp. 105–126, Apr. 2012.

[6] P. Kokol, H. B. Vošner, and J. Završnik, "Application of bibliometrics in medicine: A historical bibliometrics analysis," *Health Inf. Libraries J.*, vol. 38, no. 2, pp. 125–138, Jun. 2021.

[7] T. Mirrlees and S. Alvi, *EdTech Inc.: Selling, Automating and Globalizing Higher Education in the Digital Age*. Evanston, IL, USA: Routledge, 2019.

[8] R. Huang, J. M. Spector, and J. Yang, *Educational Technology: A Primer for the 21st Century*. Singapore: Springer, 2019.

[9] C. Jack and S. Higgins, "What is educational technology and how is it being used to support teaching and learning in the early years?" *Int. J. Early Years Educ.*, vol. 27, no. 3, pp. 222–237, Jul. 2019.

[10] A. Kirkwood and L. Price, "Technology-enhanced learning and teaching in higher education: What is 'enhanced' and how do we know? A critical literature review," *Learn., Media Technol.*, vol. 39, no. 1, pp. 6–36, Jan. 2014.

[11] D. L. Fabry and J. R. Higgs, "Barriers to the effective use of technology in education: Current status," *J. Educ. Comput. Res.*, vol. 17, no. 4, pp. 385–395, Dec. 1997.

[12] O. Scheuer and B. M. McLaren, "Educational data mining," in *Encyclopedia of the Sciences of Learning*. Boston, MA, USA: Springer, 2012.

[13] I. Atanassova, M. Bertin, and P. Mayr, "Mining scientific papers: NLP-enhanced bibliometrics," *Frontiers Res. Metrics Anal.*, vol. 4, p. 2, Apr. 2019.

[14] E. W. Hulme *et al.*, *Statistical Bibliography in Relation to the Growth of Modern Civilization*. London, U.K.: Butler & Tanner, 1923.

[15] A. Pritchard, ''Statistical bibliography or bibliometrics,'' *J. Document.*, vol. 25, no. 4, pp. 348–349, 1969.

[16] R. W. Burchfield, *A Supplement to the Oxford English Dictionary*. Oxford, U.K.: Clarendon, 1972.

[17] R. N. Broadus, ''Toward a definition of 'bibliometrics,''' *Scientometrics*, vol. 12, nos. 5–6, pp. 373–379, Nov. 1987.

[18] Y. Gao, L. Ge, S. Shi, Y. Sun, M. Liu, B. Wang, Y. Shang, J. Wu, and J. Tian, ''Global trends and future prospects of e-waste research: A bibliometric analysis,'' *Environ. Sci. Pollut. Res.*, vol. 26, no. 17, pp. 17809–17820, Jun. 2019.

[19] P. H. Lv, G.-F. Wang, Y. Wan, J. Liu, Q. Liu, and F.-C. Ma, ''Bibliometric trend analysis on global graphene research,'' *Scientometrics*, vol. 88, no. 2, pp. 399–419, Aug. 2011.

[20] S. Miau and J.-M. Yang, ''Bibliometrics-based evaluation of the blockchain research trend: 2008—March 2017,'' *Technol. Anal. Strategic Manage.*, vol. 30, no. 9, pp. 1029–1045, Sep. 2018.

[21] S. Sood, K. Rawat, and G. Sharma, ''3-D printing technologies from infancy to recent times: A scientometric review,'' *IEEE Trans. Eng. Manag.*, early access, Jan. 5, 2022, doi: 10.1109/TEM.2021.3134128.

[22] P. Buitelaar, G. Bordea, and B. Coughlan, ''Hot topics and schisms in NLP: Community and trend analysis with saffron on ACL and LREC proceedings,'' in *Proc. Lang. Resour. Eval. Conf. (LREC)*, 2014, pp. 2083–2088.

[23] A. L. Meyers, Y. He, Z. Glass, J. Ortega, S. Liao, A. Grieve-Smith, R. Grishman, and O. Babko-Malaya, ''The termolator: Terminology recognition based on chunking, statistical and search-based scores,'' *Frontiers Res. Metrics Anal.*, vol. 3, p. 19, Jun. 2018.

[24] Z. Yang, L. Hong, and B. D. Davison, ''Academic network analysis: A joint topic modeling approach,'' in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2013, pp. 324–333.

[25] G. G. Chowdhury, ''Natural language processing,'' *Annu. Rev. Inf. Sci. Technol.*, vol. 37, no. 1, pp. 51–89, 2005.

[26] E. D. Liddy, ''Natural language processing,'' in *Encyclopedia of Library and Information Science*. Great Barrier Reef, QLD, Australia: Wiley, 2001.

[27] A. F. Anta, L. N. Chiroque, P. Morere, and A. Santos, ''Sentiment analysis and topic detection of Spanish tweets: A comparative study of of NLP techniques,'' *Procesamiento del lenguaje Natural*, vol. 50, pp. 45–52, Apr. 2013.

[28] H. S. Choi, W. S. Lee, and S. Y. Sohn, ''Analyzing research trends in personal information privacy using topic modeling,'' *Comput. Secur.*, vol. 67, pp. 244–253, Jun. 2017.

[29] S. Syed and M. Spruit, ''Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation,'' in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2017, pp. 165–174.

[30] A. Abu-Jbara, J. Ezra, and D. Radev, ''Purpose and polarity of citation: Towards NLP-based bibliometrics,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2013, pp. 596–606.

[31] C. L. Streeter and D. F. Gillespie, ''Social network analysis,'' *J. Social Service Res.*, vol. 16, nos. 1–2, pp. 201–222, 1993.

[32] J. Scott, ''Social network analysis,'' *Sociology*, vol. 22, no. 1, pp. 109–127, 1988.

[33] E. Otte and R. Rousseau, ''Social network analysis: A powerful strategy, also for the information sciences,'' *J. Inf. Sci.*, vol. 28, no. 6, pp. 441–453, 2002.

[34] A. Marin and B. Wellman, ''Social network analysis: An introduction,'' in *The SAGE Handbook of Social Network Analysis*, vol. 11. London, U.K.: Sage, 2011, p. 25.

[35] G. González-Alcaide, J. Park, C. Huamaní, J. Gascón, and J. M. Ramos, ''Scientific authorships and collaboration network analysis on Chagas disease: Papers indexed in PubMed (1940–2009),'' *Revista do Instituto de Medicina Tropical de São Paulo*, vol. 54, no. 4, pp. 219–228, Aug. 2012.

[36] S. Dawson, D. Gašević, G. Siemens, and S. Joksimovic, ''Current state and future trends: A citation network analysis of the learning analytics field,'' in *Proc. 4th Int. Conf. Learn. Anal. Knowl.*, Mar. 2014, pp. 231–240.

[37] Y. Kajikawa and Y. Takeda, ''Citation network analysis of organic LEDs,'' *Technol. Forecasting Soc. Change*, vol. 76, no. 8, pp. 1115–1123, 2009.

[38] K. Harman, A. Koohang, and J. Paliszkiewicz, ''Scholarly interest in gamification: A citation network analysis,'' *Ind. Manage. Data Syst.*, vol. 114, no. 9, pp. 1438–1452, Oct. 2014.

[39] Y. Kajikawa, J. Ohno, Y. Takeda, K. Matsushima, and H. Komiyama, ''Creating an academic landscape of sustainability science: An analysis of the citation network,'' *Sustainability Sci.*, vol. 2, no. 2, pp. 221–231, Oct. 2007.

[40] R. Huang, J. M. Spector, and J. Yang, ''Introduction to educational technology,'' in *Educational Technology*. Singapore: Springer, 2019, pp. 3–31.

[41] M. Bond, O. Zawacki-Richter, and M. Nichols, ''Revisiting five decades of educational technology research: A content and authorship analysis of the British journal of educational technology,'' *Brit. J. Educ. Technol.*, vol. 50, no. 1, pp. 12–63, Jan. 2019.

[42] X. Chen, G. Yu, G. Cheng, and T. Hao, ''Research topics, author profiles, and collaboration networks in the top-ranked journal on educational technology over the past 40 years: A bibliometric analysis,'' *J. Comput. Educ.*, vol. 6, no. 4, pp. 563–585, Dec. 2019.

[43] F. Gurcan, O. Ozyurt, and N. E. Cagitay, ''Investigation of emerging trends in the e-learning field using latent Dirichlet allocation,'' *Int. Rev. Res. Open Distrib. Learn.*, vol. 22, no. 2, pp. 1–18, Jan. 2021.

[44] M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. García. (2021). *Analyzing Trends and Patterns Across the Educational Technology Communities Using Fontana Framework*. [Online]. Available: https://osf.io/nr425/?view_only=2e712bc2217044938e41f6321890d688

[45] Elsevier. (2021). *About Scopus*. [Online]. Available: https://www.elsevier.com/es-es/solutions/scopus

[46] (2021). *Web of Science*. [Online]. Available: https://apps.webofknowledge.com/

[47] J. A. Palmer. (2021). *pdftotext*. Accessed: Oct. 5, 2021. [Online]. Available: https://pypi.org/project/pdftotext/

[48] L. Yao, D. Mimno, and A. McCallum, ''Efficient methods for topic model inference on streaming document collections,'' in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 937–946.

[49] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, ''An analysis of the coherence of descriptors in topic modeling,'' *Expert Syst. Appl.*, vol. 42, no. 13, pp. 5645–5657, Aug. 2015.

[50] M. Röder, A. Both, and A. Hinneburg, ''Exploring the space of topic coherence measures,'' in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, Feb. 2015, pp. 399–408.

[51] S. Kapadia. (2019). *Evaluate Topic Models: Latent Dirichlet Allocation (LDA)*. [Online]. Available: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[52] S. Kumar, ''Co-authorship networks: A review of the literature,'' *Aslib J. Inf. Manage.*, vol. 67, no. 1, pp. 55–73, 2015.

[53] W. Lu, J. Janssen, E. Milios, N. Japkowicz, and Y. Zhang, ''Node similarity in the citation graph,'' *Knowl. Inf. Syst.*, vol. 11, no. 1, pp. 105–129, 2007.

[54] NetworkX. (2021). *NetworkX—Network Analysis in Python*. Accessed: Jun. 14, 2021. [Online]. Available: https://networkx.org

[55] PyVis. (2021). *PyVis Introduction*. Accessed: Jun. 14, 2021. [Online]. Available: https://pyvis.readthedocs.io/en/latest/introduction.html

[56] IAALDE. (2021). *International Alliance to Advance Learning in the Digital Era (IAALDE)*. [Online]. Available: https://alliancelss.com

[57] M. Bastian, S. Heymann, and M. Jacomy, ''Gephi: An open source software for exploring and manipulating networks,'' in *Proc. 3rd Int. AAAI Conf. Weblogs Social Media*, 2009, pp. 361–362.

[58] L.-K. Lee, S. K. S. Cheung, and L.-F. Kwok, ''Learning analytics: Current trends and innovative practices,'' *J. Comput. Educ.*, vol. 7, no. 1, pp. 1–6, 2020.

[59] S. Kross and P. J. Guo, ''Students, systems, and interactions: Synthesizing the first four years of learning@scale and charting the future,'' in *Proc. 5th Annu. ACM Conf. Learn. Scale*, Jun. 2018, pp. 1–10.

[60] P. Blikstein, ''Multimodal learning analytics,'' in *Proc. 3rd Int. Conf. Learn. Anal. Knowl.*, 2013, pp. 102–106.

[61] R. F. Kizilcec, C. Piech, and E. Schneider, ''Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses,'' in *Proc. 3rd Int. Conf. Learn. Anal. Knowl.*, 2013, pp. 170–179.

[62] X. Chen, D. Zou, and H. Xie, ''Fifty years of British journal of educational technology: A topic modeling based bibliometric perspective,'' *Brit. J. Educ. Technol.*, vol. 51, no. 3, pp. 692–708, May 2020.

[63] J.-L. Hung, ''Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics,'' *Brit. J. Educ. Technol.*, vol. 43, no. 1, pp. 5–16, Jan. 2012.

[64] F. Fidler and J. Wilcox, ''Reproducibility of scientific results,'' in *The Stanford Encyclopedia of Philosophy*. Stanford, CA, USA: Metaphysics Research Lab, Stanford Univ., 2018.

[65] T. van der Zee and J. Reich, ''Open education science,'' *AERA Open*, vol. 4, no. 3, 2018, Art. no. 2332858418787466.

[66] S. Buck, ''Solving reproducibility,'' *Science*, vol. 348, no. 6242, pp. 1403–1403, 2015. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aac8041, doi: 10.1126/science.aac8041.

MANUEL J. GOMEZ received the B.Sc. degree with a focus on applied computing and data science and the M.Sc. degree in big data. He is currently pursuing the Ph.D. degree in computer science with the University of Murcia, Spain. He is a member of the CyberDataLab, University of Murcia. His research interests include data mining, educational technology, game-based assessment, and natural language processing.

JOSÉ A. RUIPÉREZ-VALIENTE (Senior Member, IEEE) received the B.Eng. degree in telecommunications from the Universidad Católica de San Antonio de Murcia and the M.Eng. degree in telecommunications and the M.Sc. and Ph.D. degrees in telematics from the Universidad Carlos III of Madrid while conducting research with the Institute IMDEA Networks in the area of learning analytics and educational data mining. He was a Postdoctoral Associate with MIT. He has received more than 20 academic/research awards and fellowships, has published more than 90 scientific publications in high impact venues, and participated in over 18 funded projects. He currently holds the prestigious Spanish Fellowship Juan de la Cierva with the University of Murcia.

FÉLIX J. GARCÍA CLEMENTE received the M.Sc. and Ph.D. degrees in computer science from the University of Murcia, Spain. He is currently an Associate Professor at the Department of Computer Engineering, UMU. His teaching include courses in computer networks, network management, ubiquitous computing, and mobile device programming. His major research interests include cybersecurity, distributed management of networks and services, and interaction systems.