PIXEL BIT REVISTA DE MEDIOS Y EDUCACIÓN

Píxel-BIT Revista de Medios y Educación - 2020 - nº 57

ISSN: 1133-8482. e-ISSN: 2171-7966

Big Data Irruption in Education

Irrupción del Big Data en la Educación

Dr. Antonio Matas Terrón amatas@uma.es

Dr. Juan José Leiva Olivencia jjleiva@uma.es

Dr. Pablo Daniel Franco Caballero pablo.franco@uma.es

Universidad de Málaga. Facultad de Educación. Departamentos de: Teoría e Historia de la Educación; Métodos de Investigación y Diagnóstico en Educación; Didáctica y Organización Escolar. Bulevar Louis Pasteur, 25, 29010 Málaga (España).

ABSTRACT

The objective is to analyse the production of scientific articles on Big Data in Education from 2013 to 2018, as well as to identify the most frequently used keywords in those articles. The publications of the Scopus database were consulted using a search algorithm based on pre-established criteria. Through a quantitative procedure, including text mining, different aspects of the production of research articles on Big Data in Education were analysed: citations, authors, journals, and topics covered.

The results show an increase in production over Big Data in Education from 2015, as well as a change in trend in the subjects dealt with, going from studies focused on Psychology and Behaviour to studies focused on Education. From this point, there is a real interest in this field of research, and the usage in the Educational System will change the pedagogical mentality and in the training centres.

KEYWORDS

Big data; text mining; education; natural language process; technology.

RESUMEN

El objetivo de este documento es analizar la producción de artículos científicos sobre "Big Data" en Educación desde 2013 hasta 2018, además de identificar las palabras clave más frecuentes en esos artículos. Para ello, se consultaron publicaciones en la base de datos Scopus usando un algoritmo de búsqueda basado en un criterio preestablecido. A través de un proceso cuantitativo, incluyendo la minería de textos, fueron analizados diferentes aspectos de la producción de artículos de investigación sobre Big Data en Educación: citas, autores, revistas y temas fueron considerados.

Los resultados muestran un incremento de producción sobre Big Data en Educación desde 2015, al igual que un cambio en la tendencia de los temas tratados, partiendo de estudios enfocados en Psicología y comportamiento, llegando a estudios más enfocados en Educación. De esto se deduce un gran interés en este campo de investigación y su aplicación en el Sistema Educativo, que cambiará su mentalidad pedagógica, además de la de los propios centros formativos.

PALABRAS CLAVE

Big data; minería de textos; educación; procesamiento de lenguaje natural; tecnología.

Recibido: 08-05-2019 | Revisado: 27-05-2019 | Aceptado: 04-10-2019 | Publicado: 01-01-2020
DOI: https://doi.org/10.12795/pixelbit.2020.i57.02 | Páginas 59-90

1.- Introduction

Education is being continuously affected by several "revolutionary changes", like communication, movies and television, computers, internet, virtual reality, and artificial intelligence. However, the high expectations generated by Information and Communication Technologies (ICTs) and the digitization of human processes and acts cannot be understood without a more human, global and inclusive social and educational perspective. This is not only a question of including technology in the classroom, but understanding it within the framework of a change that may provoke an improvement in the teaching and learning process (Fosso Wamba, Akter, Edwards, Chopin, & Gnanzou, 2015). The better we understand how the student interacts with educational technologies, the better we can optimize the teaching-learning processes.

As these educational technologies collect big data about their use, they can be used to analyse the educational process (Reyes, 2015). For this reason, we have seen a growing trend in the number of papers published about Big Data in Education. This trend means that the professionals of Education have noticed the potential of using massive data in Education. It hardly means that Big Data could be considered as the core of an educational revolution, but it could become a new paradigm of Education. This paradigm will help to understand how to plan, develop and improve the Educational process. Anyway, the use of Big Data in Education means not only several opportunities but also problems that we have to face both critically and ethically (Williamson, 2017).

Big Data is not simply related to the quantity, but also to the relation between data (Fosso Wamba *et al.*, 2015). The concept of Big Data applies to all information that can't be processed or analysed using traditional tools or processes; it is focused on the data-driven decision making (Puyol, 2014). Its value comes from all the connections and patterns that can be established between the pieces of information gathered by all the devices that surround the individual: the location, the visited websites, the time spent on each one, the group of acquaintances, how their information is structured, how they interact with the smart device applications, etc. (Boyd & Crawford, 2011). The literature talks about seven V's that define Big Data: volume, velocity, variety, veracity, validity, volatility and value (Mayer-Schönberger & Cukier, 2018). The goal of Big Data is to include the information from different sources both for the business process, the organization of an enterprise, and for other institutions. Educators can take advantage of

the "value" of the information gathered from each digital educational resource, and the variety of data available, as soon as they are being used to establish connections and patterns to improve the educational process, adapting it to the learner.

Research in Education continues focusing on analysing what can be measured and, once measured; establish relation between the gathered data. Academic results, teacher and satisfaction surveys on the process experienced at a specific point are still being gathered for this purpose (Mayer-Schönberger & Cukier, 2018). During the last years several researches have been carried out that emphasize the need to experiment and analyse Big Data in Educational centres (Anshari, Alas, & Guan, 2016; Dede, 2016; Hilbert, 2016; Long & Siemens, 2011; Waller & Fawcett, 2013). Managing digitized data on student performance, traits, and learning styles can help teachers to improve the design, development, and evaluation of learning processes. Using Big Data, the approach could evolve to gathering information during the development of the training process, using ICTs to continuously gather all the information: from the time spent on each task, tools used, contacts, places and ways of execution, timetables, etc.

The didactical dimension is related to the existence or non-existence of technological resources that allow the intense and innovative use of ICTs for all students, generating searches for relevant patterns of knowledge about their learning processes (Arranz & Alonso, 2013; Mayer-Schönberger & Cukier, 2018). Here it is fundamental to have digital devices and to promote didactic strategies that allow the revision and valuation of the data stored in them. As an example, the Massive Open Online Courses (MOOCs) are training and Educational interaction sites that generate large amounts of information about students' behaviour and evolution (Khan, Uddin, & Gupta, 2014).

In order to design more effective and personalized teaching, it is essential to evaluate the actions of students in a digital environment, without falling into reductionist didactic parameters (Reyes, 2015). It is obvious that there is a positive correlation between student activity and participation and academic success. However, some studies raise the difficulty that analyses based on massive data have in providing coherent and solid pedagogical information (Zablith, 2015).

The intersectional dimension in the understanding of the use of Big Data in education implies the recognition of the necessary improvements from an integral methodological perspective. We mean that the use of Big Data requires the implementation of Learning Analytics and, therefore, the incorporation

of new active and innovative methodologies (Ellaway, Pusic, Galbraith, & Cameron, 2014). There is another term that is also related to data analysis applied to big data, which is Academic Analytics. To clarify the difference between Learning Analytics and Academic Analysis, we can say that Learning Analytics is the process of collecting educational information, measuring, analysing and reporting these findings from the student's learning process and its context in order to understand his or her learning process. Academic analysis, on the other hand, applies business intelligence tools for decision making in educational institutions; its objective is, therefore, to improve the educational institution by collecting, measuring and generating reports that identify the strengths and weaknesses of the institution.

The future looks promising; however, the question is if the field of education has sufficient interest in it. This question can be answered analysing the bibliographic production about Big Data in Education.

Taking into account the framework drawn previously, the aim of this research is to quantify the bibliographic production of articles about Big Data from 2013 to 2018 in order to know if the presence of Big Data in Education is increasing. To do this, we analyse the total number of articles in Scopus database, the number of citations received from the main journals that published these articles and most prolific authors in this issue using a text mining based methodology (Aria & Cuccurullo, 2017). Besides, a second aim is proposed, that is to classify the production knowing their main topics.

2.- Methodology

This study follows the methodology used in similar (Crossland et *al.*, 2019; Laude, 2017). In order to retrieve the sources, the study has used the database Scopus for its recognition and international impact (Delgado-López-Cózar & Repiso-Caballero, 2013). A quantification of the bibliographic data of the usual articles in this type of studies (Chiu & Ho, 2005; Davis & Gonzalez, 2003) was developed: authors, countries, institutions and thematic areas. The frequencies of the appearance of the keywords were also quantified (Viedma-Del-Jesus, Perakakis, Muñoz, López-Herrera, & Vila, 2011).

For the search, the following criteria was used:

- ((TITLE-ABS-KEY (bigdata) OR TITLE-ABS-KEY (big AND data) AND TITLE-ABS-KEY (Education)).

- From 2013 to 2018.
- Document type: conference paper & article.
- Field: Social Science.

The indicators analysed were:

- The total number of articles per year.
- Authors.
- Production by country.
- Production magazines.
- Methodological approach.
- Issues based on keywords.

The initial information obtained a result of 832 documents, out of which 118 are available for open access. For the analysis, Scopus automated resources, as well as Orange Data Mining version 3.18 (Demšar *et al*., 2013), were used.

To reach the second goal, Natural Language Process (NLP) techniques were developed. This techniques process text with computers in order to analyse it, to extract information and eventually to represent the same information differently (Rao & McMahan, 2019). For this purpose, Orange software was used. It is a multiplatform data mining software developed in Python with text mining modules included. On the other hand, Orange has a graphical interface that allows a very intuitive programming. As of January 2019, Orange is available free of charge on the website https://orange.biolab.si.

With regard to text data, it is important to know that their techniques or methods work on different depth levels:

- Identifying words that belong to words lists (bag of words).
- Identifying of word chains as sentences or expressions.

- Identifying semantic elements, that is, detect words according to the meaning.

NLP includes many techniques that are selected according to the study aim. Those techniques can be ordered depending on their depth level. For example, tokenisation split text into tokens and it is useful to determine the analysis unit. Nevertheless, tokenisation is usually used as an initial technique to develop other analysis later (Silge & Robinson, 2017).

In this study, the following techniques were conducted among others during the pre-processing:

- Tokenization using regular expression (words). At the same time, inputs were transformed in data applying lower case and removing url's options.
- Filtering by a "Stopwords" implemented in Orange.
- Then, a bag of words was generated from input corpus. The parameter included counting the number of occurrences of every word in every document (row) and then, applied the Inverse Document Frequency (Laude, 2017).
- Next, in order to categorize the documents, a Ward linkage in Hierarchical Clustering was conducted. Previously, the Euclidean distance was calculated among others options, as Cosin distance or Manhattan distance.

The analysis design used in Orange, based in the Natural Language Process (NLP) is shown in Figure 1.

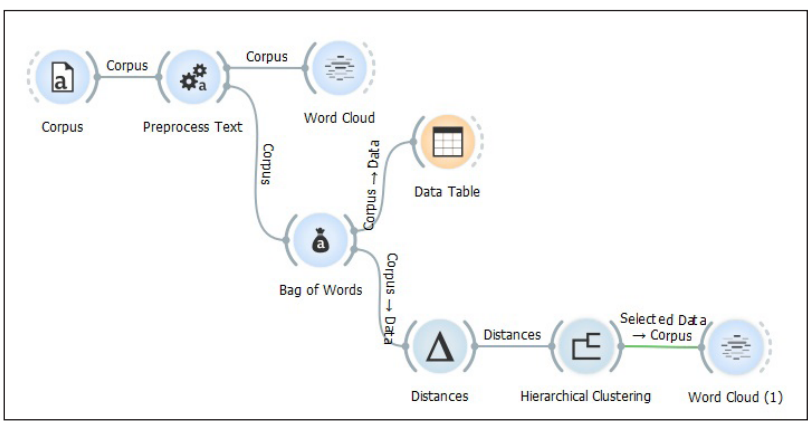

Figure 1. Analysis design in Orange

3.- Results

As shown in Table 1, the number of articles in Scopus is 832. The largest number of publications are concentrated in 2016 although the highest increase in production was in 2014. The difference between years indicates the number of papers that exceed or fall in relation to the previous year.

Table 1. Number of documents per year

Year	Frequency	Difference between years	% Difference
2013	58	-	-
2014	89	31	53%
2015	133	44	49%
2016	194	61	46%
2017	187	-7	-4%
2018	171	-16	-9%

Regarding the number of citations, there is a notable difference between articles, not only because of the year of publication, but because of the scientific impact of some articles (e.g., (Kyriakidis, Happee, & de Winter, 2015)). There are 29 documents with more than 10 cites. Table 2 shows the frequencies of the number of cites along the years.

Table 2. Number of cites per year

Number of Cities	< 2014	2014	2015	2016	2017	2018
0	46	34	19	5		1
1		2	11	4	1	1
2	1	5	2	8	3	2
3		2	8	6	4	5
4		1	1	5	9	4
5		1		6	5	3
6		1	1	2	7	9
7			3	1	6	3
8		1	1		1	4
9				1	1	2
10				2		2
>10			1	7	10	11

No publications from 2019 have been considered

The max number of total cites is for Kyriakidis' (2015) paper with 101 followed by (Graham & Shelton, 2013)'s (2013) paper with 77. The average number of citations per year since the year of publi-cation of the article has been calculated. Subsequently, those documents with an average higher than 10 are shown in Table 3. On the other hand, the USA is the country with more articles (232) fol-lowed by China (142) and United Kingdom (69).

Transportation Research Part F is the journal with more cites. On the other hand, most of the documents are published as Journal papers (536). However, there is also a high number of Conference Proceedings (294). The rest are book publications. The first publication language is English (774). The second language is Spanish with only 18 documents and third are Chinese and Portuguese, each one with 9 documents.

Table 3. Authors of articles which the number of cites is higher than ten

Autors	Average	Journal title
Kyriakidis M., Happee R., De Winter J.C.F.	25.25	Transportation Research Part F: Traffic Psychology and Behaviour
Graham M., Shelton T.	15.4	Dialogues in Human Geography
Daniel B.	17.75	British Journal of Educational Technology
Williamson B.	18.66	Journal of Education Policy
Selwyn N.	13.25	Learning, Media and Technology
Khan M.A.-U.-D., Uddin M.F., Gupta N.	10	Journal of Vocational Behaviour

The total number of keywords was 3735. The diversity of terms was very broad, which generated a low frequency of repetition. Table 4 offers only those with a frequency higher than 10. The word Big Data (and synonyms) is the term that is most frequently repeated, as expected. Second is Learning Analytics. Education and Higher Education are also frequently repeated.

Table 4. Frequency of keywords

Keyword	Frequency
Big data	161
Learning analytics	42
Education	41
Higher Education	35
MOOCs	19
Data mining	16
Personality	11

The most used separate words as keywords are “Data” (374 times) “Education” (308 times) “Learn-ing” (228 times) “Analytics” (102 times) and of course “Big” (232 times). The rest of words are used less than 100 times from the whole keyword camp. Text mining was developed on the keywords, not only on sparse words but also on whole keyword expressions (Figure 1). A Word Preprocess was applied on the 834 documents, involving 7444 tokens and 2173 total types. Besides, the analysis was developed counting the frequency of terms and applying an inverse document frequency algorithm with a normalization based on the sum of elements.

A Euclidean distance metric was selected in order to develop a hierarchical cluster. The cluster was developed computing the distance between the clusters’ most distant elements. To facilitate the in-terpretation, only four levels of max depth of pruning were selected. The dendrogram from the anal-ysis is in Figure 2. Two clusters are clear at 75% height ratio. As is seen, the first cluster includes documents published in 2013 and 2014 and the second cluster includes documents published from 2015 to 2018.

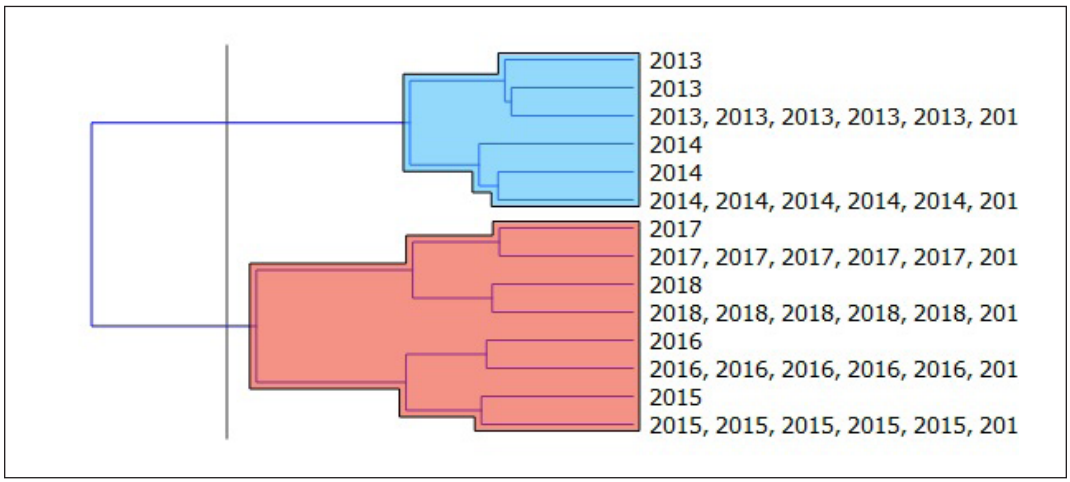

Figure 2. Hierarchical clusters of the document keywords by publication year

The first cluster (years 2013 and 2014) has 147 documents with 629 words. The higher frequency words are Education (48), Data (45), Learning (30) and Big (26). The rest of the words have less than 15 references. Figure 3 shows the word cloud, where text size is directly related to frequency.

In the second cluster (from 2015 to 2018) includes 687 documents with 1920 words. Those with a frequency higher than 100 are Data (329), Education (260), Big (206) and Learning (198). Figure 4 shows the word cloud.

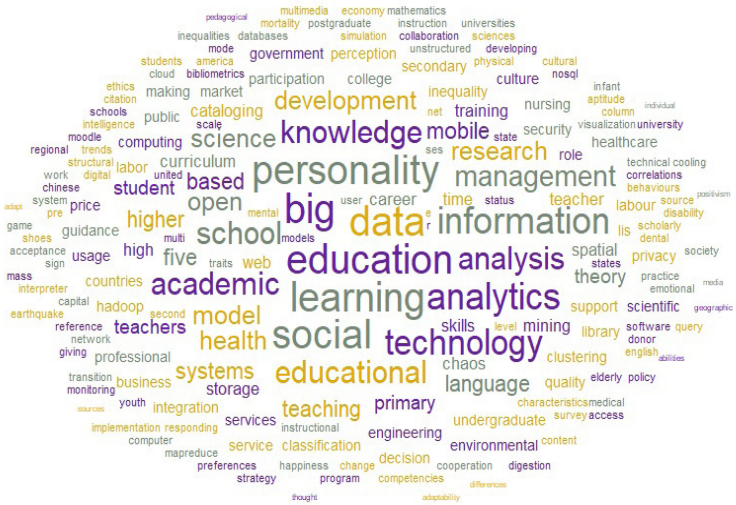

Figure 3. Word

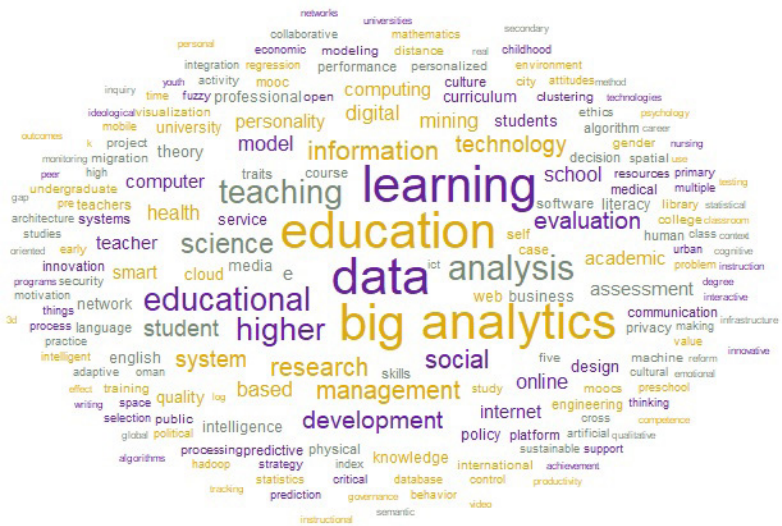

Figure 4. Word cloud from the keyword cluster two (2015-2018 publications)

4.- Discussion

Results show that there is a real interest in this field of research. Nevertheless, it is important to es-tablish previously some framework that helps to avoid "apophenia", it means, finding connections between unrelated themes (Boyd & Crawford, 2011).

In relation to the number of papers (Table 1), up to 2016 there is an increase in the number of articles (average of 50%). Since then, the number of articles published has fallen slightly (average of -6%). Therefore, it is necessary to analyse if an isolated peak occurred and, in the future, the topic will lose interest, or we are dealing with a stabilization of production. On the other hand, our hypothesis is this it is only the beginning of a stable production period.

In the results, it is observed that some specific articles systematically have more citations. This means that some articles are becoming references within Big Data in Education. It is estimated that these works will reach the status of classical works, for example, the paper of Kyriakidis, Happee and Winter (2015). On the other hand, the journal with the highest number of citations received about this subject is "Transportation Research Part F: Traffic Psychology and Behaviour". Therefore, Big Data arises in Education linked to the Psychology of traffic above all. However, the subject has been quickly integrated into the field of general Education.

The cluster analysis reveals two clear groups based on the year of publication. In both periods, the themes that stand out are Education, Big Data and Learning. However, the first group highlights the topic of Psychology (personality above all) as observed in the word cloud. The second cluster focuses on Educational and social themes. In general, these results are consistent with the topics treated by authors such as Ellaway *et al.* (2014).

In Education, Big Data is linked to University Education above all. Some authors state that Big Data has appeared in Education in recent years thanks to technological advances, as well as the development of e-learning, especially Massive Open Online Courses (MOOCs) (Mayer-Schönberger & Cukier, 2018). However, the topic of MOOC's does not stand out from others in the results of this study.

Big Data means a change in the paradigm of content organization. A transcendental element in the understanding of the irruption of Big Data in Education is linked to the need to generate new peda-

gogical proposals and ideas. The idea is to make an optimal use of all sources of data about students' performance and behaviour, and Big Data in Education should preferably be translated into more personalised training (Ferguson & Shum, 2012; Williamson, 2016). The main challenge could be the possibility to exchange data between analysis software, so a flexible and consistent semantic model that can handle both new and existing sources is needed (Modoni, Doukas, Terkaj, Sacco, & Mourtzis, 2017).

This may lead both universities and schools to change their structure, including their material resources. It will also affect the architecture of their buildings (Rathore, Ahmad, Paul, & Rho, 2016) or cities, due to the Internet of Things (IoT), where all devices capable of interconnecting and communicating with each other with Internet (vacuums, lights, windows, air conditioners, etc.). The digitization of Education is an emerging reality that requires the design and development of operational and effective proposals and initiatives (Chen, Mao, & Liu, 2014).

The use of Big Data in the Educational System implies a change in pedagogical mentality and in the very institutional conception of training centres (Borgman, 2015). We must investigate the possibilities of combined or blended curriculum design and development, which means accepting that school organizations are no longer privileged spaces for socialization and personalization of learning in terms of current configuration (Macfadyen, Dawson, Pardo, & Gaševic, 2014). The promotion of mixed Educational models, including face-to-face and online classes with mentoring and coaching, are being promoted from higher Education. It is not risky to state that sooner rather than later there will be a proliferation of combined models of adaptive learning in all post-compulsory Education, including high school and vocational training, and even in Compulsory Secondary Education.

The point no longer lies in the emergence of Big Data in the Educational field, but in the responsible, innovative and inclusive use of technological tools to activate operational mechanisms for the processing and analysis of massive data (Crossley, 2014). The organizational and professional restructuring of Educational centres involves promoting an academic culture favourable to generating learn-ing analytics, implementing concrete experiences of innovation and didactic change. This means abandoning the uncritical collection of data and the lack of motivation caused by the uselessness of complex information collection processes without any kind of feedback, transfer or application in everyday classroom practice (Gašević, Dawson, & Siemens, 2015).

Daniel, (2019) sets the main difference between educational research with big data instead of without big data: the context of the researchers might be unknown; the emergent epistemology and ontology, instead of the focused one; might require real-time analysis; the need of web mining applications, sensors and traffic monitoring instead of standalone software, etc. We agree the need of an ontological orientation because, especially at qualitative research, where the data collection is critical, because big data analysis might use data already collected, so it might need some kind of validation (Boyd & Crawford, 2012). Related to this, Big Data researches have been found about measurement of emotional states, collecting the communication elements related to their interventions in social networks; that is why ethical implications and the ontological approach of the data will be a priority (Conway & O'Connor, 2016; George, Haas, & Pentland, 2014).

All this could also be applied to the business context. We must not forget the paradigm of the "teaching factory" that integrates industry and academia through ICT in a context of continuous education (Chryssolouris, Mavrikios, & Rentzos, 2016). This paradigm has to use Big Data as the key to constantly improve the learning process analysing the time spent, the failures or success at the tasks and the times that searches for more information to personalize the e-learning.

4.- Conclusions

Following the review of these documents, it has become clear that Big Data is having a growing impact on education, not only in the field of research, but also in practice by making use of Learning Management Systems (LMS).

We are witnessing a digital revolution with Educational consequences and repercussions of huge potential. Its increasing visibility will generate opportunities for personalized and adaptive learning, especially if we are able to take an intelligent and strategic step from the systematic and coherent collection of information to its effective analysis.

Social, cultural and economic digitization is unstoppable, and the Educational world cannot remain on the side-lines of these incessant processes of change, transformation and technological innovation. It does not seem reasonable that some schools are stuck in the past, not providing opportunities for success for

all students, excluding the learning of digital skills and, in the worst case, not analysing in a reflective and critical way the databases about their performance, learning styles and personal training needs.

We must underline the need to stimulate initial and ongoing teacher training in understanding the benefits and pedagogical potential of Big Data. The development of specific training objectives in terms of learning analytics is a key element for the construction of personalized learning that is adapted and redefined in real time according to the needs and progress detected through the inclusive analysis of pedagogical data.

Therefore, we suspect that future lines of research about Big Data in Education will be:

- To analyse implications for universities in the future: role of teachers, managers and the organization of universities. It is important, due to the need to adapt the contents, the teaching methodology, the different activities focused on improve the specific skills and, obviously, the implication of the teaching-learning process related to the professional world.
- To explore new ways to evaluate students, teachers, institutions and curriculum content.
- To analyse the ethical consequences that Big Data implies in relation to privacy.
- To analyse the legal consequences.
- To study the effects that the individualization of learning implies in the socialization of the student.

Irrupción del Big Data en la Educación

1.- Introducción

La educación se ve afectada continuamente por varios "cambios revolucionarios", como la comunicación, el cine y televisión, los ordenadores, Internet, la realidad virtual y la inteligencia artificial. Sin embargo, las grandes expectativas generadas por las Tecnologías de la Información y la Comunicación (TIC) y la digitalización de los procesos y actos humanos no pueden entenderse sin una perspectiva social y educativa más humana, global e inclusiva. No se trata sólo de incluir la tecnología en el aula, sino de entenderla en el marco de un cambio que pueda provocar una mejora en el proceso de enseñanza y aprendizaje (Fosso Wamba, Akter, Edwards, Chopin, & Gnanzou, 2015). Cuanto mejor entendamos cómo interactúa el estudiante con las tecnologías educativas, mejor podremos optimizar los procesos de enseñanza-aprendizaje.

Como estas tecnologías educativas recogen grandes datos sobre su uso, pueden ser utilizadas para analizar el proceso educativo (Reyes, 2015). Por esta razón, hemos visto una tendencia creciente en el número de artículos publicados sobre Big Data en Educación. Esta tendencia significa que los profesionales de la Educación han notado el potencial de usar esos datos masivos en la Educación. Esto difícilmente significa que el Big Data pueda ser considerado como el núcleo de una revolución educativa, pero esto podría convertirse en un nuevo paradigma de la Educación. Este paradigma ayudará a entender cómo planificar, desarrollar y mejorar el proceso educativo. De todos modos, el uso del Big Data en la educación no sólo implica varias oportunidades, sino también problemas a los que tenemos que enfrentarnos de manera crítica y ética (Williamson, 2017).

El Big Data no se relaciona únicamente con la cantidad de datos, sino también con la relación entre ellos (Fosso Wamba *et al.*, 2015). El concepto de Big Data se aplica a toda la información que no puede ser procesada o analizada utilizando herramientas o procesos tradicionales; se centra, mayormente, en la toma de decisiones basadas en evidencias o datos (Puyol, 2014). Su valor proviene de todas las conexiones y patrones que se pueden establecer entre las informaciones recogidas por todos los dispositivos que rodean al individuo: la ubicación, los sitios web visitados, el tiempo dedicado a cada uno de ellos, el grupo de conocidos, cómo se estructura su información, cómo interactúan con las aplicaciones de los

dispositivos inteligentes, etc. (Boyd & Crawford, 2011). La literatura habla de siete V's que definen a Big Data: volumen, velocidad, variedad, veracidad, validez, volatilidad y valor (Mayer-Schönberger & Cukier, 2018). El objetivo de Big Data es incluir la información de diferentes fuentes tanto para procesos y líneas de negocio o de organización de una empresa, como para otras instituciones. Los educadores pueden aprovechar el "valor" de la información recogida de cada recurso educativo digital, y la variedad de datos disponibles, desde el momento en que se utilizan para establecer conexiones y patrones para mejorar el proceso educativo, adaptándolo al alumno.

La investigación en Educación sigue centrándose en analizar lo que se puede medir y, una vez medido, establecer una relación entre los datos recogidos. Los resultados académicos, el profesorado y las encuestas de satisfacción sobre el proceso educativo experimentado en un momento determinado siguen siendo recogidos para este fin (Mayer-Schönberger & Cukier, 2018). Durante los últimos años se han llevado a cabo varias investigaciones que enfatizan la necesidad de experimentar y analizar Big Data en los centros educativos (Anshari, Alas, & Guan, 2016; Dede, 2016; Hilbert, 2016; Long & Siemens, 2011; Waller & Fawcett, 2013). La gestión de datos digitalizados sobre el rendimiento, los rasgos y los estilos de aprendizaje de los estudiantes pueden ayudar a los profesores a mejorar el diseño, el desarrollo y la evaluación de los procesos de aprendizaje. Utilizando Big Data, el enfoque podría evolucionar hacia la recogida de información durante el desarrollo del proceso de formación, utilizando las TIC para recoger continuamente toda la información: desde el tiempo dedicado a cada tarea, las herramientas utilizadas, los contactos, los lugares y formas de ejecución, los calendarios, etc.

La dimensión didáctica está relacionada con la existencia o no de recursos tecnológicos que permitan el uso intenso e innovador de las TIC para todo el alumnado, generando búsquedas de patrones relevantes de información sobre sus procesos de aprendizaje (Arranz & Alonso, 2013; Mayer-Schönberger & Cukier, 2018). Aquí es fundamental contar con dispositivos digitales y promover estrategias didácticas que permitan la revisión y valoración de los datos almacenados en ellos. Por ejemplo, los Cursos Abiertos Masivos en Línea (MOOCs) son sitios de formación e interacción educativa que generan gran cantidad de información sobre el comportamiento y la evolución de los estudiantes (Khan, Uddin & Gupta, 2014).

Para diseñar una enseñanza más eficaz y personalizada, es fundamental evaluar las acciones de los alumnos en un entorno digital, sin caer en parámetros didácticos reduccionistas (Reyes, 2015). Es obvio que existe una correlación positiva entre la actividad y participación de los estudiantes con el éxito académico. Sin embargo, algunos estudios plantean la dificultad que tienen los análisis basados en Big Data para proporcionar información pedagógica coherente y sólida (Zablith, 2015).

La dimensión transversal en la comprensión del uso del Big Data en la Educación implica el reconocimiento de las mejoras necesarias desde una perspectiva metodológica integral. Nos referimos a que el uso de Big Data requiere la implementación de Analíticas del Aprendizaje (Learning Analytics) y, por lo tanto, la incorporación de nuevas metodologías activas e innovadoras (Ellaway, Pusic, Galbraith, & Cameron, 2014). Hay otro término que también está relacionado con el análisis de datos en Big Data, que es el Análisis Académico. Para aclarar la diferencia entre la Analítica de Aprendizaje y el Análisis Académico, podemos decir que la Analítica de Aprendizaje es el proceso de recolectar información educativa, medirla, analizarla y reportar los hallazgos en el proceso de aprendizaje del estudiante y su contexto, con el fin de entender su proceso de aprendizaje. El Análisis Académico, por su parte, aplica herramientas de inteligencia de negocios para la toma de decisiones en las instituciones educativas; su objetivo es, por lo tanto, mejorar la institución educativa mediante la recolección, medición y generación de informes que identifiquen las fortalezas y debilidades de la institución.

El futuro parece prometedor; sin embargo, la cuestión es si el entorno educativo tiene suficiente interés en ello. Esta pregunta puede ser respondida analizando la producción bibliográfica sobre Big Data en Educación.

Teniendo en cuenta el marco trazado anteriormente, el objetivo de esta investigación es cuantificar la producción bibliográfica de artículos sobre Big Data de 2013 a 2018 para saber si la presencia de Big Data en la educación está aumentando. Para ello, analizamos el número total de artículos en la base de datos de Scopus, el número de citas recibidas de las principales revistas que publicaron estos artículos y, finalmente, los autores más prolíficos en este ámbito utilizando una metodología basada en la minería de textos (Aria & Cuccurullo, 2017). Además, se propone un segundo objetivo, que es clasificar dicha producción científica en función de sus temas principales.

2.- Metodología

Este estudio sigue la metodología utilizada en otros estudios similares (Crossland et *al.*, 2019; Laude, 2017). Para acceder a las fuentes, el estudio ha utilizado la base de datos Scopus debido a su reconocimiento e impacto internacional (Delgado-López-Cózar & Repiso-Caballero, 2013). Se desarrolló una cuantificación de los datos bibliográficos de los artículos habituales en este tipo de estudios (Chiu & Ho, 2005; Davis & González, 2003): autores, países, instituciones y áreas temáticas. También se cuantificaron las frecuencias de aparición de las palabras clave (Viedma-Del-Jesús, Perakakis, Muñoz, López-Herrera, & Vila, 2011).

Para la búsqueda, se usó el siguiente criterio:

- Palabras clave en el título: ([“bigdata” O (“big” Y “data”)] Y “Education”).
- Desde 2013 hasta 2018.
- Tipo de documentos: Artículos y actas conferencias.
- Campo: Ciencias Sociales.

Los indicadores analizados fueron:

- Número total de artículos por año.
- Autores.
- Producción según país.
- Revistas y monográficos producidos.
- Aproximación metodológica.
- Temáticas basadas en las palabras clave.

La información inicial obtuvo un resultado de 832 documentos, de los cuales 118 están disponibles en acceso abierto. Para el análisis se utilizaron recursos automatizados de Scopus, así como el programa Orange Data Mining versión 3.18 (Demšar *et al.*, 2013).

Para alcanzar el segundo objetivo, se desarrollaron técnicas de Procesamiento de Lenguaje Natural (PLN). Estas técnicas procesan textos mediante ordenadores para analizarlo, extraer información y, finalmente, representar la información de manera diferente (Rao & McMahan, 2019). Para ello, se utilizó el programa Orange. Es un software de minería de datos multiplataforma desarrollado en Python que incluye módulos para la minería de textos. Por otro lado, Orange dispone de una interfaz gráfica que permite una programación muy intuitiva. En enero de 2019, Orange está disponible gratuitamente en el sitio web https://orange.biolab.si

En relación a los datos de texto, es importante saber que sus técnicas o métodos funcionan en diferentes niveles de profundidad:

- Identificar palabras que pertenecen a listas de palabras (bolsa de palabras).
- Identificación de cadenas de palabras como frases o expresiones.
- Identificar elementos semánticos, es decir, detectar palabras según su significado.

El PLN incluye muchas técnicas que se seleccionan de acuerdo con el objetivo del estudio. Estas técnicas pueden ser ordenadas dependiendo de su nivel de profundidad. Por ejemplo, la señalización (tokenization) divide el texto en señales (tokens) y es muy útil para determinar la unidad de análisis. Sin embargo, la señalización se suele utilizar como una técnica inicial para desarrollar otros análisis (Silge & Robinson, 2017).

En este estudio se realizaron, entre otras, las siguientes técnicas durante el pre-procesamiento:

- Señalización usando expresiones regulares (palabras). Al mismo tiempo, los datos de entrada fueron transformados en minúsculas y eliminando las opciones de hiperenlaces.
- Filtrado por un "Palabras reservadas" ("stopwords") implementado en Orange.
- Luego, se generó una bolsa de palabras a partir del bloque de entradas. El parámetro incluía contar el número de ocurrencias de cada palabra en cada documento (fila) y luego, aplicar la Frecuencia Inversa de Documento (Inverse Document Frequency) (Laude, 2017).
- A continuación, con el fin de categorizar los documentos, se llevó a cabo un análisis jerárquico

usando el algoritmo de Ward. Previamente, la distancia euclídea fue calculaba frente a otras opciones, como la distancia Cosin o la distancia Manhattan.

El diseño de análisis utilizado en Orange, basado en el Proceso de Lenguaje Natural (PLN) se muestra en la Figura 1.

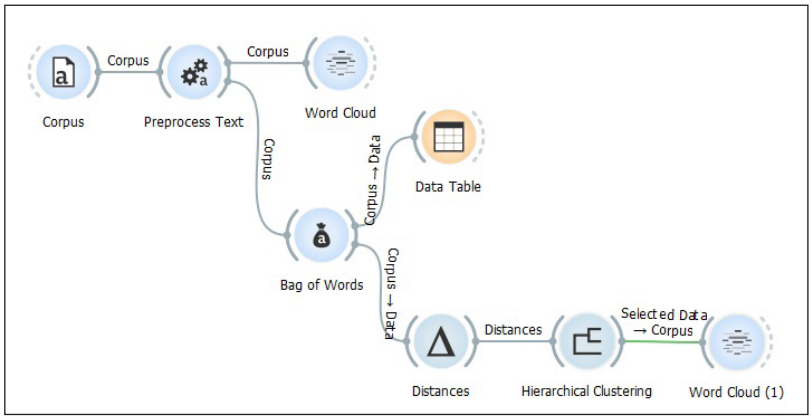

Figura 1. Diseño del análisis en Orange

3.- Resultados

Como se muestra en la Tabla 1, el número de artículos en Scopus es de 832. El mayor número de publicaciones se concentra en 2016, aunque el mayor aumento proporcional de la producción se produjo en 2014. La diferencia entre años indica el número de artículos que exceden o decaen en relación al año anterior.

Tabla 1. Número de documentos por año

Año	Frecuencia	Diferencia entre años	% Diferencia
2013	58	-	-
2014	89	31	53%
2015	133	44	49%
2016	194	61	46%
2017	187	-7	-4%
2018	171	-16	-9%

En cuanto al número de citas, hay una diferencia notable entre los artículos, no sólo por el año de publicación, sino por el impacto científico de algunos de ellos (por ejemplo, Kyriakidis, Happee, & de Winter,

2015). Hay 29 documentos con más de 10 citas. La Tabla 2 muestra las frecuencias del número de citas a lo largo de los años.

Tabla 2. Número de citas por año

Número de citas por año	< 2014	2014	2015	2016	2017	2018
0	46	34	19	5		1
1		2	11	4	1	1
2	1	5	2	8	3	2
3		2	8	6	4	5
4		1	1	5	9	4
5		1		6	5	3
6		1	1	2	7	9
7			3	1	6	3
8		1	1		1	4
9				1	1	2
10				2		2
>10			1	7	10	11

No se consideraron las publicaciones de 2019

El documento más citado es el de Kyriakidis et al., (2015) con 101 citas, seguido del papel de Graham & Shelton (2013) con 77. Se ha calculado el número medio de citaciones por año desde el año de publicación del artículo. Los documentos con un promedio superior a 10 se muestran en la Tabla 3. Por otro lado, Estados Unidos es el país con más artículos publicados (232), seguido de China (142) y el Reino Unido (69).

"Transportation Research Part F" es la revista más citada. Por otra parte, la mayoría de los documentos se publican en forma de artículos de revista (536). Sin embargo, también hay un gran número de Actas de conferencias (294). El resto son publicaciones de libros. El idioma principal de publicación es el inglés (774). El segundo idioma más usado es el español con sólo 18 documentos y el tercer idioma más utilizado corresponde tanto al chino como al portugués, cada uno con 9 documentos.

El número total de palabras clave fue de 3735. La diversidad de términos era muy amplia, lo que generaba una baja frecuencia de repetición. La Tabla 4 ofrece sólo aquellos términos con una frecuencia superior a 10. La palabra Big Data (y sinónimos) es el término que se repite con más frecuencia, como se esperaba. La segunda palabra clave es Analíticas de Aprendizaje ("Learning Analytics"). Educación y Educación Superior también se repiten con frecuencia.

Tabla 3. Autores de artículos con más de 10 citas

Autores	Media de citas por año	Título de la revista
Kyriakidis M., Happee R., De Winter J.C.F.	25.25	Transportation Research Part F: Traffic Psychology and Behaviour
Graham M., Shelton T.	15.4	Dialogues in Human Geography
Daniel B.	17.75	British Journal of Educational Technology
Williamson B.	18.66	Journal of Education Policy
Selwyn N.	13.25	Learning, Media and Technology
Khan M.A.-U.-D., Uddin M.F., Gupta N.	10	Journal of Vocational Behaviour

Tabla 4. Frecuencia de palabras clave

Palabras clave	Frecuencia
Big data	161
Learning analytics	42
Education	41
Higher Education	35
MOOCs	19
Data mining	16
Personality	11

Las palabras clave individuales más utilizadas son "Datos" (374 veces) "Educación" (308 veces) "Aprendizaje" (228 veces) "Analíticas" (102 veces) y, por supuesto, "Big" (232 veces). El resto de palabras se repiten menos de 100 veces en el total. La minería de textos se llevó a cabo sobre las palabras clave, no sólo sobre palabras individuales sino también sobre expresiones completas (Figura 1). A los 834 documentos se les aplicó un pre-procesamiento de palabras, con 7444 señales (tokens) y 2173 tipos en total. Además, el análisis se desarrolló contabilizando la frecuencia de los términos y aplicando un algoritmo de frecuencia inversa de documento (idf) con una normalización basada en la suma de los elementos.

Se seleccionó una métrica de distancia euclídea para elaborar el conglomerado (cluster) jerárquico. El agrupamiento por conglomerados fue desarrollado calculando la distancia entre los elementos más distantes entre sí. Para facilitar la interpretación, sólo se seleccionaron cuatro niveles de profundidad de poda máxima. El dendrograma del análisis realizado puede observarse en la Figura 2. Dos agrupaciones están claramente definidas al 75% de profundidad. Como se puede ver, el primer grupo incluye los

documentos publicados en 2013 y 2014 y en el segundo grupo, se incluyen documentos publicados entre 2015 y 2018.

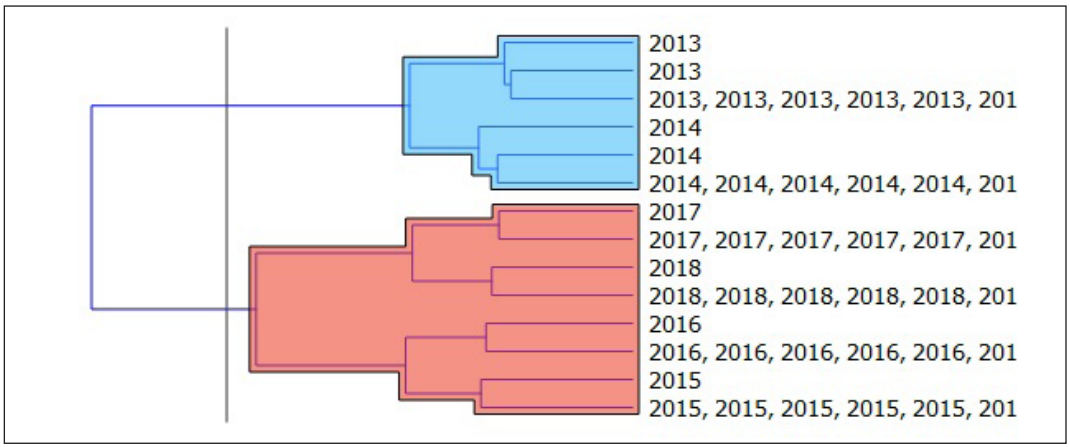

Figura 2. Agrupaciones jerárquicas de las palabras clave según el año de publicación

El primer grupo (años 2013 y 2014) tiene 147 documentos con 629 palabras. Las palabras de mayor frecuencia son Educación (48), Datos (45), Aprendizaje (30) y Big (26). El resto de las palabras tienen menos de 15 referencias. La Figura 3 muestra la nube de palabras, donde el tamaño del texto está directamente relacionado con la frecuencia.

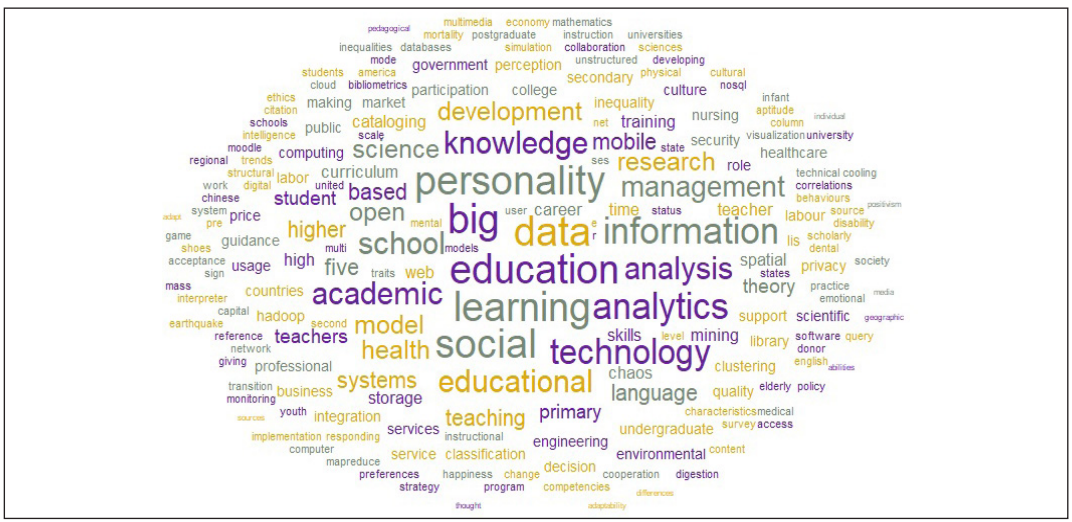

Figura 3. Nube de palabras de la primera agrupación (publicaciones de años 2013-2014)

En el segundo grupo (de 2015 a 2018) incluye 687 documentos con 1920 palabras. Los que tienen una frecuencia superior a 100 son "Datos" (329), "Educación (260), Big (206) y "Aprendizaje" (198). La Figura 4 muestra la palabra nube.

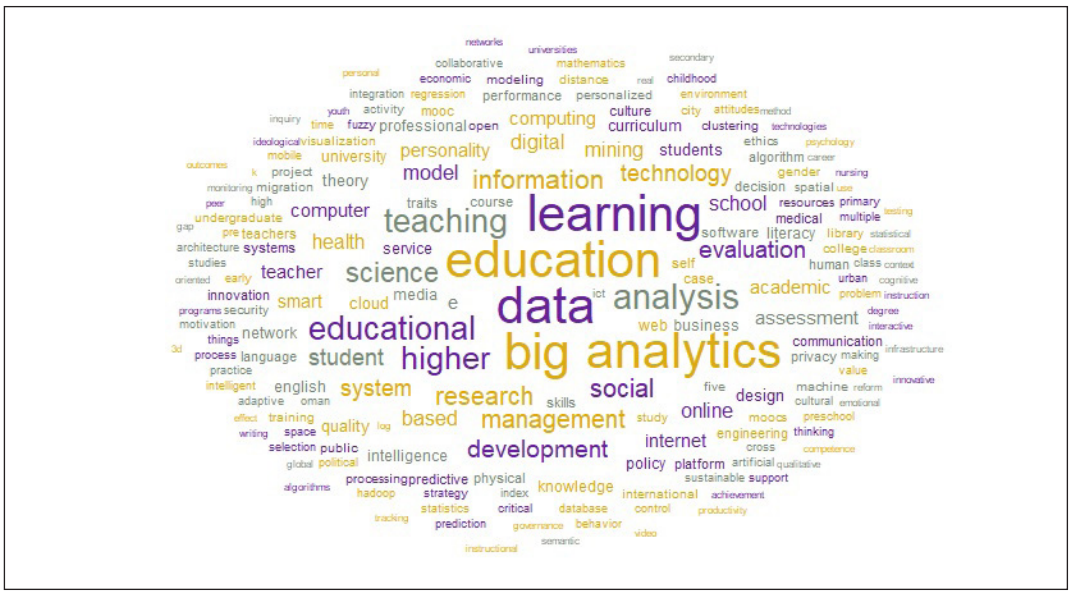

Figure 4. Nube de palabras de la segunda agrupación (publicaciones de años 2015-2018)

4.- Discusión

Los resultados muestran que existe un interés real en este campo de investigación. Sin embargo, es importante establecer previamente algún marco que ayude a evitar la "apofenia", es decir, encontrar conexiones entre temas no relacionados (Boyd & Crawford, 2011).

En relación con el número de artículos (Tabla 1), hasta 2016 hay un aumento en el número de artículos (una media del 50%). Desde entonces, el número de artículos publicados ha disminuido ligeramente (media del -6%). Por lo tanto, es necesario analizar si se produjo un pico aislado y, en el futuro, el tema perderá interés, o si se trata de una estabilización de la producción. Por otro lado, nuestra hipótesis es que sólo es el comienzo de un período de producción estable.

En los resultados, se observa que algunos artículos específicos tienen sistemáticamente más citas. Esto significa que algunos artículos se están convirtiendo en referencias dentro de Big Data en la Educación.

Se estima que estas obras alcanzarán el estatus de obras clásicas, un ejemplo de ello, es el artículo de Kyriakidis, Happee y Winter (2015). Por otro lado, la revista con mayor número de citas recibidas sobre este tema es "Transportation Research Part F: Traffic Psychology and Behaviour". Por lo tanto, Big Data surge en la Educación relacionada, sobre todo, a la Psicología del transporte. Sin embargo, el tema se ha integrado rápidamente en el campo de la educación general.

El análisis de conglomerados revela dos grupos claros basados en el año de publicación 2015. En ambos períodos, los temas que se destacan son Educación, Big Data y Aprendizaje. Sin embargo, el primer grupo destaca el tema de la Psicología (la personalidad, sobre todo) como se observa en la nube de palabras. El segundo grupo se centra en temas educativos y sociales. En general, estos resul-tados son coherentes con los temas tratados por autores como Ellaway *et al.*, (2014).

En Educación, Big Data está vinculado sobre todo a la Educación Universitaria. Algunos autores afirman que el Big Data ha aparecido en educación en los últimos años gracias a los avances tecnológicos, así como al desarrollo del e-learning, especialmente los Cursos Masivos Abiertos en Línea (MOOCs) (Mayer-Schönberger & Cukier, 2018). Sin embargo, el tema de los MOOCs no destaca significativamente del resto en los resultados de este estudio.

El uso de Big Data significa un cambio en el paradigma de la organización de contenidos. Un elemento trascendental en la comprensión de la irrupción del Big Data en la Educación está ligado a la necesidad de generar nuevas propuestas e ideas pedagógicas. La idea es hacer un uso óptimo de todas las fuentes de datos acerca del rendimiento y el comportamiento de los estudiantes, y el Big Data, en relación con la educación, debería traducirse esencialmente en una formación más personalizada (Ferguson & Shum, 2012; Williamson, 2016). El principal desafío podría ser la posibilidad de intercambiar datos entre programas orientados al análisis, por lo que se necesita un modelo semántico flexible y coherente que pueda manejar tanto las fuentes de información nuevas como las ya existentes (Modoni, Doukas, Terkaj, Sacco, & Mourtzis, 2017).

Esto puede llevar tanto a las universidades como a las escuelas a cambiar su estructura, incluidos sus recursos materiales. También afectará a la arquitectura de sus edificios (Rathore, Ahmad, Paul, & Rho, 2016) o ciudades, debido a la Internet de las Cosas (IoT), donde todos los dispositivos son capaces de interconectarse y comunicarse entre sí mediante Internet (aspiradoras, luces, ventanas, aparatos de aire

acondicionado, etc.). La digitalización de la Educación es una realidad emergente que requiere el diseño y desarrollo de propuestas e iniciativas operativas y efectivas (Chen, Mao y Liu, 2014).

El uso del Big Data en el Sistema Educativo implica un cambio en la mentalidad pedagógica y en la propia concepción institucional de los centros de formación (Borgman, 2015). Debemos investigar las posibilidades de diseño y desarrollo de currículos combinados o flexibles, lo que significa aceptar que las organizaciones escolares ya no son espacios privilegiados para la socialización y personalización del aprendizaje en términos de la configuración actual (Macfadyen, Dawson, Pardo, & Gaševic, 2014). Desde la Educación Superior se está promoviendo la promoción de modelos educativos mixtos, incluyendo clases presenciales y en línea unidas a tutorías y asesoramientos (coaching). No es arriesgado afirmar que, tarde o temprano, habrá una proliferación de modelos combinados de aprendizaje adaptativo en toda la educación postobligatoria, incluyendo la formación superior, la formación profesional, e incluso en la educación secundaria obligatoria.

Ya no tratamos la aparición del Big Data en el campo de la educación, sino de un uso responsable, innovador e inclusivo de herramientas tecnológicas para activar mecanismos operativos de procesamiento y análisis de datos masivos (Crossley, 2014). La reestructuración organizativa y profesional de los centros educativos implica la promoción de una cultura académica favorable a la creación de analíticas de aprendizaje, y a la implementación de experiencias concretas de innovación y cambio didáctico. Esto significa abandonar la recolección acrítica de datos y la falta de motivación causada por la inutilidad de procesos complejos de recolección de información sin ningún tipo de retroalimentación, transferencia o aplicación en la práctica diaria del aula (Gašević, Dawson y Siemens, 2015).

Daniel (2019) establece la principal diferencia entre el uso (o no) de Big Data en la investigación educativa: el contexto de los investigadores podría ser desconocido; la epistemología y ontología emergente, en lugar de la actual; podría requerir de análisis en tiempo real; la necesidad de aplicaciones de minería web, sensores y monitoreo de tráfico en lugar de software autónomo, etc. Estamos de acuerdo en la necesidad de una orientación ontológica, especialmente en la investigación cualitativa donde la recolección de datos es crítica, porque los análisis de Big Data podrían usar datos ya recolectados, que podrían necesitar algún tipo de validación (Boyd & Crawford, 2012). En relación a esto, se han encontrado investigaciones de Big Data sobre la medición de los estados emocionales, recogiendo los elementos de comunicación

relacionados a las intervenciones en las redes sociales; es por eso que las implicaciones éticas y el enfoque ontológico de los datos serán una prioridad (Conway & O'Connor, 2016; George, Haas, & Pentland, 2014).

Todo esto también podría aplicarse al contexto empresarial. No debemos olvidar el paradigma de la "fábrica de enseñanza" que integra industria y academia a través de las TIC en un contexto de educa-ción continua (Chryssolouris, Mavrikios, & Rentzos, 2016). Este paradigma podría utilizar Big Data como elemento clave para mejorar constantemente el proceso de aprendizaje analizando el tiempo empleado, los fracasos o éxitos en las tareas, y el tiempo invertido en buscar información adicional, para personalizar el e-learning.

5.- Conclusiones

Tras la revisión de estos documentos, ha quedado claro que el Big Data está teniendo un impacto creciente en la educación, no sólo en el campo de la investigación, sino también en la práctica, al hacer uso de los Sistemas de Gestión del Aprendizaje (LMS).

Estamos asistiendo a una revolución digital con consecuencias y repercusiones educativas de gran potencial. Su creciente visibilidad generará oportunidades de aprendizaje personalizado y adaptativo, especialmente si somos capaces de dar un paso inteligente y estratégico desde la recolección sistemática y coherente de la información hasta su análisis efectivo.

La digitalización social, cultural y económica es imparable, y el mundo de la educación no puede permanecer al margen de estos incesantes procesos de cambio, transformación e innovación tecnológica. No parece razonable que algunas escuelas estén estancadas en el pasado, sin ofrecer oportunidades para el éxito de todos los estudiantes, excluyendo el aprendizaje de habilidades digitales y, en el peor de los casos, que no analicen de forma reflexiva y crítica las bases de datos sobre su rendimiento, estilos de aprendizaje y necesidades de formación personal.

Debemos subrayar la necesidad de estimular la formación inicial y continua de los profesores para que comprendan los beneficios y el potencial pedagógico del Big Data. El desarrollo de objetivos formativos específicos en términos de analítica del aprendizaje es un elemento clave para la construcción de un

aprendizaje personalizado, adaptado y redefinido, en tiempo real, en función de las necesidades y avances detectados a través del análisis inclusivo de los datos pedagógicos.

Por lo tanto, sospechamos que las futuras líneas de investigación sobre Big Data en Educación estarán relacionadas con:

- Analizar las implicaciones para las universidades en el futuro: el papel de los profesores, los gestores y la organización de las universidades. Es importante, debido a la necesidad de adaptar los contenidos, la metodología de enseñanza, las diferentes actividades enfocadas a mejorar las habilidades específicas y, obviamente, la implicación del proceso de enseñanza-aprendizaje relacionado con el mundo profesional.
- Explorar nuevas formas de evaluar a los estudiantes, profesores, instituciones y contenidos curriculares.
- Analizar las consecuencias éticas que implica el uso de Big Data en relación con la privacidad.
- Analizar las consecuencias legales.
- Estudiar los efectos que la individualización del aprendizaje implica en la socialización del alumno

Referencias

Anshari, M., Alas, Y., & Guan, L. S. (2016). Developing online learning resources: Big data, social networks, and cloud computing to support pervasive knowledge. *Education and Information Technologies, 21*(6), 1663-1677. https://doi.org/10.1007/s10639-015-9407-3

Aria, M., & Cuccurullo, C. (2017). bibliometrix : An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Arranz, O., & Alonso, V. (2013). *Big Data & Learning Analytics: A Potential Way to Optimize eLearning Technological Tools.* Recuperado de https://bit.ly/2VweSsw7

Borgman, C. L. (2015). *Big data, little data, no data: scholarship in the networked world.* Cambridge, Massachusetts: The MIT Press.

Boyd, D., & Crawford, K. (2011). Six Provocations for Big Data. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.1926431

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society, 15*(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications, 19*(2), 171-209. https://doi.org/10.1007/s11036-013-0489-0

Chiu, W.-T., & Ho, Y.-S. (2005). Bibliometric analysis of homeopathy research during the period of 1991 to 2003. *Scientometrics, 63*(1), 3-23. https://doi.org/10.1007/s11192-005-0201-7

Chryssolouris, G., Mavrikios, D., & Rentzos, L. (2016). The Teaching Factory: A Manufacturing Education Paradigm. *Procedia CIRP, 57*, 44-48. https://doi.org/10.1016/j.procir.2016.11.009

Conway, M., & O'Connor, D. (2016). Social media, big data, and mental health: current advances and ethical implications. *Current Opinion in Psychology, 9*, 77-82. https://doi.org/10.1016/j.copsyc.2016.01.004

Crossland, T., Stenetorp, P., Riedel, S., Kawata, D., Kitching, T. D., & Croft, R. A. C. (2019). Towards Machine-assisted Meta-Studies: The Hubble Constant. *arXiv:1902.00027 [astro-ph].* Recuperado de https://bit.ly/2Ow7YC1

Crossley, M. (2014). Global league tables, big data and the international transfer of educational research modalities. *Comparative Education, 50*(1), 15-26. https://doi.org/10.1080/03050068.2013.871438

Daniel, B. K. (2019). Big Data and data science: A critical review of issues for educational research: Critical issues for educational research. *British Journal of Educational Technology, 50*(1), 101-113. https://doi.org/10.1111/bjet.12595

Davis, J. C., & Gonzalez, J. G. (2003). Scholarly Journal Articles about the Asian Tiger Economies: authors, journals and research fields, 1986-2001. *Asian-Pacific Economic Literature, 17*(2), 51-61. https://doi.org/10.1046/j.1467-8411.2003.00131.x

Dede, C. J. (2016). Next steps for "Big Data" in education: Utilizing data-intensive research. *Educational*

Technology. Recuperado de https://bit.ly/31ZiEwK

Delgado-López-Cózar, E., & Repiso-Caballero, R. (2013). The Impact of Scientific Journals of Communication: Comparing Google Scholar Metrics, Web of Science and Scopus. *Comunicar, 21*(41), 45-52. https://doi.org/10.3916/C41-2013-04

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., … Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research, 14*, 2349-2353.

Ellaway, R. H., Pusic, M. V., Galbraith, R. M., & Cameron, T. (2014). Developing the role of big data and analytics in health professional education. *Medical Teacher, 36*(3), 216-222. https://doi.org/10.3109/0142159X.2014.874553

Ferguson, R., & Shum, S. B. (2012). Social learning analytics: five approaches. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, 23. https://doi.org/10.1145/2330601.2330616

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics, 165*, 234-246. https://doi.org/10.1016/j.ijpe.2014.12.031

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64-71. https://doi.org/10.1007/s11528-014-0822-x

George, G., Haas, M. R., & Pentland, A. (2014). Big Data and Management. *Academy of Management Journal, 57*(2), 321-326. https://doi.org/10.5465/amj.2014.4002

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography, 3*(3), 255-261. https://doi.org/10.1177/2043820613513121

Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges. *Development Policy Review, 34*(1), 135-174. https://doi.org/10.1111/dpr.12142

Khan, M. A., Uddin, M. F., & Gupta, N. (2014). Seven V's of Big Data understanding Big Data to extract value. *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*, 1-5.

https://doi.org/10.1109/ASEEZone1.2014.6820689

Kyriakidis, M., Happee, R., & de Winter, J. C. F. (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation Research Part F: Traffic Psychology and Behaviour, 32*, 127-140. https://doi.org/10.1016/j.trf.2015.04.014

Laude, H. (2017). *Data scientist y lenguaje R: guía de autoformación para el uso de Big Data* (F. J. Piqueres Juan, Trad.). Barcelona: Eni.

Long, P., & Siemens, G. (2011, septiembre 12). Penetrating the Fog: Analytics in Learning and Education. *EduCause Review, 46*(5), 6.

Macfadyen, L. P., Dawson, S., Pardo, A., & Gaševic, D. (2014). Embracing Big Data in Complex Educational Systems: The Learning Analytics Imperative and the Policy Challenge. *Research & Practice in Assessment, 9,* 17-28.

Mayer-Schönberger, V., & Cukier, K. (2018). *Aprender con big data.* Madrid: Turner.

Modoni, G. E., Doukas, M., Terkaj, W., Sacco, M., & Mourtzis, D. (2017). Enhancing factory data integration through the development of an ontology: from the reference models reuse to the semantic conversion of the legacy models. *International Journal of Computer Integrated Manufacturing, 30*(10), 1043-1059. https://doi.org/10.1080/0951192X.2016.1268720

Puyol, J. (2014). Una aproximación a Big Data. *Revista de Derecho de la UNED (RDUNED), 14*, 471-506. https://doi.org/10.5944/rduned.14.2014.13303

Rao, D., & McMahan, B. (2019). *Natural language processing with PyTorch: build intelligent language applications using deep learning.* Sebastopol, CA: OReilly Media.

Rathore, M. M., Ahmad, A., Paul, A., & Rho, S. (2016). Urban planning and building smart cities based on the Internet of Things using Big Data analytics. *Computer Networks, 101*, 63-80. https://doi.org/10.1016/j.comnet.2015.12.023

Reyes, J. A. (2015). The skinny on big data in education: Learning analytics simplified. *TechTrends, 59*(2), 75-80. https://doi.org/10.1007/s11528-015-0842-1

Silge, J., & Robinson, D. (2017). *Text mining with R: a tidy approach*. Beijing; Boston: O'Reilly.

Viedma-Del-Jesus, M. I., Perakakis, P., Muñoz, M. Á., López-Herrera, A. G., & Vila, J. (2011). Sketching the first 45 years of the journal Psychophysiology (1964-2008): A co-word-based analysis: Forty-five years of Psychophysiology. *Psychophysiology, 48*(8), 1029-1036. https://doi.org/10.1111/j.1469-8986.2011.01171.x

Waller, M. A., & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics, 34*(2), 77-84. https://doi.org/10.1111/jbl.12010

Williamson, B. (2016). Digital education governance: data visualization, predictive analytics, and 'real-time' policy instruments. *Journal of Education Policy, 31*(2), 123-141. https://doi.org/10.1080/02680939.2015.1035758

Williamson, B. (2017). *Big data in education: the digital future of learning, policy and practice*. Thousand Oaks, CA: SAGE Publications.

Zablith, F. (2015). Interconnecting and Enriching Higher Education Programs Using Linked Data. *Proceedings of the 24th International Conference on World Wide Web*, 711–716. https://doi.org/10.1145/2740908.2741740

Cómo citar este artículo:
Franco Caballero, P. D., Matas Terrón, A. & Leiva Olivencia, J. J. (2020). Big Data Irruption in Education. [Irrupción del Big Data en la Educación]. *Pixel-Bit. Revista de Medios y Educación,* 57, 59-90. https://doi.org/10.12795/pixelbit.2020.i57.02