





Article

# Machine Learning and Big Data in the Impact Literature. A Bibliometric Review with Scientific Mapping in Web of Science

Jesús López Belmonte <sup>1</sup>, Adrián Segura-Robles <sup>2,\*</sup>, Antonio-José Moreno-Guerrero <sup>1</sup> and María Elena Parra-González <sup>2</sup>

<sup>1</sup> Department of Didactics and School Organization, University of Granada, 51001 Ceuta, Spain; [jesuslopez@ugr.es](mailto:jesuslopez@ugr.es) (J.L.B.); [ajmoreno@ugr.es](mailto:ajmoreno@ugr.es) (A.-J.M.-G.)

<sup>2</sup> Department of Research Methods and Diagnosis in Education, University of Granada, 51001 Ceuta, Spain; [elenaparra@ugr.es](mailto:elenaparra@ugr.es)

\* Correspondence: [adrianseg@ugr.es](mailto:adrianseg@ugr.es)

Received: 20 February 2020; Accepted: 16 March 2020; Published: 27 March 2020



**Abstract:** Combined use of machine learning and large data allows us to analyze data and find explanatory models that would not be possible with traditional techniques, which is basic within the principles of symmetry. The present study focuses on the analysis of the scientific production and performance of the Machine Learning and Big Data (MLBD) concepts. A bibliometric methodology of scientific mapping has been used, based on processes of estimation, quantification, analytical tracking, and evaluation of scientific research. A total of 4240 scientific publications from the Web of Science (WoS) have been analyzed. Our results show a constant and ascending evolution of the scientific production on MLBD, 2018 and 2019 being the most productive years. The productions are mainly in English language. The topics are variable in the different periods analyzed, where “machine-learning” is the one that shows the greatest bibliometric indicators, it is found in most of motor topics and is the one that offers the greatest line of continuity between the different periods. It can be concluded that research on MLBD is of interest and relevance to the scientific community, which focuses its studies on the branch of machine-learning.

**Keywords:** scientific production; bibliometric analysis; machine learning; big data; web of science

## 1. Introduction

The idea of Machine Learning was not unique in computing, but due to the consistently varying nature of necessities of the present world it has come up in unique forms. With the expansion of the web, a large amount of advanced data are being created, which implies that there are much more data accessible for machines to learn and analyze. Today, calculations of machine learning empower the computers to speak with autonomously driven cars and humans, compose and coordinate reports, and find accused terrorists. Supervised, unsupervised and reinforcement learning are the three sub-areas of Machine Learning [1].

Some Machine Learning techniques for processing of Big Data are not efficient and are not adaptable to get together a high volume, value, velocity, and variety, hence it requests to rehash itself for handling of big data [2]. Adaptability is a difficult problem with conventional calculations of machine learning [3]. In the event that a machine learning approach is utilized to address a calculation deficiency and a material science-based model is accessible, at that point numerical outcomes might be adequate in requests to process acceptable execution measures [2]. Machine learning is utilized in Web search, spam channels, advertisement situation, recommender frameworks, credit scoring,

stock exchanging, and misrepresentation recognition, tranquilizer structure, and numerous different applications [4].

Everybody when thinking of machine learning thinks about overfitting, yet it comes in numerous structures that are not promptly self-evident. Machine learning utilized on Big Data has an extraordinary potential for a forecast [5]. Big Data examination is one of the extraordinary difficulties for Machine Learning calculations because most real-life applications include an enormous data or big data information base [6]. In the advanced world, data are created from different sources and the rapid change from progressive improvements has encouraged the growth of big data [7,8]. It gives developing accomplishments in many fields with an assortment of huge datasets. All data accessible in the type of big data are not helpful for examination or dynamic process. Industry and the scholarly world are keen on spreading the discoveries of big data [9].

Machine learning approaches incorporate choice tree learning, support vector machines (SVMs), Bayesian networks, artificial neural systems, clustering, and hereditary calculations, and so forth [10]. Its target is to find information and settle on making intelligent choices and decisions. Deep machine learning is a machine learning strategy, where numerous layers of data handling stages are exploited in various leveled designs [11,12]. It figures various leveled highlights or portrayals of the observational data, where the more elevated level highlights are characterized by lower-level ones. Deep learning algorithms extricate significant level, complex deliberations as data portrayals through various leveled learning process [13,14].

People are experiencing an incredible innovative progression and an ever-increasing need to get data since it is authoritative. Machine learning has been used in Big Data, and in numerous sets of fields [15–17]. It includes an assortment of devices, strategies, and advancements for processing on data gainfully, at any measure. Increment in the limit of storage and innovations in progressive storage improved preparing limit of present-day computers and accessibility of huge scope data—all prompts the improvement in the processing field of Big Data [18]. Current occasion's equipment and programming advances can manage, operate process and break down a humongous measure of data as at no other time. Very huge arrangements of information can be gathered and investigated to uncover examples, designs, and affiliations identified with human conduct and interaction. Everybody is preparing Big Data, and attempting to get the advantages by processing and utilizing different Big Data handling systems [6].

Conventional social databases are not prepared to take care of large data [19]. To direct these datasets is problematic with the conventional data getting ready structures [20,21]. Besides, data stockpiling, data perception, data progress, data infiltrating, data security and examination, data protection infringement and sharing propose distinctive tough difficulties that the Big Data strengthens [22]. It is almost pervasive. In each business, for example, wellbeing or general expectations for everyday comforts could apply large data analytics [23]. This can be practiced by expertly picturing and exhibiting the data in a reasonable way.

Big Data Analytics is a rising edge for movement and advancement of developments. It gives data-equal usage of scientific, measurable and machine-learning calculations for organized and unstructured data. The issues in Big Data are not many and keeping in mind that receiving the innovation skillfully, one ought to obviously considered by its association [16,24]. The importance of this research is that other [25,26] studies usually focus on a series of bibliometric indicators such as authorship index, nationality of authors, university of origin, language, number of references per year or the number of downloads per article. This research uses the bibliometrics to focus on the analysis of co-words and bibliometric indicators such as the h-index, thus generating maps with nodes, which can show the performance and location of several conceptual subdomains related to the terms “Machine Learning” and “Big Data”.

## 2. Purpose of Study

In the present study the concepts “Machine Learning” and “Big Data” (MLBD) in the scientific literature registered in the Web of Science (WoS) database are analyzed. For the development of this research, scientific mapping will be used based on the measurement of different bibliometric indicators and the dynamic and structural development of the delimited constructs. Previous studies of impact journals of the Journal Citation Reports (JCR) have been taken as a methodological model with the purpose of following a research method validated by specialists in this type of analysis [27,28].

The purpose of this study is to analyze the path and projection of both terms in the indexed publications in the main WoS collection. First, the database was analyzed to inquire about the state of the matter and verify the existence of studies that have treated the concepts presented at the bibliometric level. As a result, no study was reported in which MLBD were related and analyzed using the scientific mapping technique.

This work assumes an exploratory component that contributes to the reduction of the gap produced in the literature found in WoS. In this line, the findings reached here will be a breakthrough in science by presenting new results that may arouse the interest of other researchers to continue studying in this state of the art.

The objectives proposed in this research are:

To know the performance of scientific production indexed in WoS alluding to MLBD.

To determine the scientific evolution of MLBD in WoS.

To create the most incidents about MLBD in WoS.

To find out the most influential authors in MLBD in WoS.

## 3. Materials and Methods

### 3.1. Research Design

Bibliometrics was the methodology used to develop the study and achieve the scope of the proposed objectives. The choice of this research approach was based on the greatness of Scientometrics for the search, registration, analysis and prediction of scientific literature [29]. For optimal development, the guidelines of experts in bibliometric were followed [30].

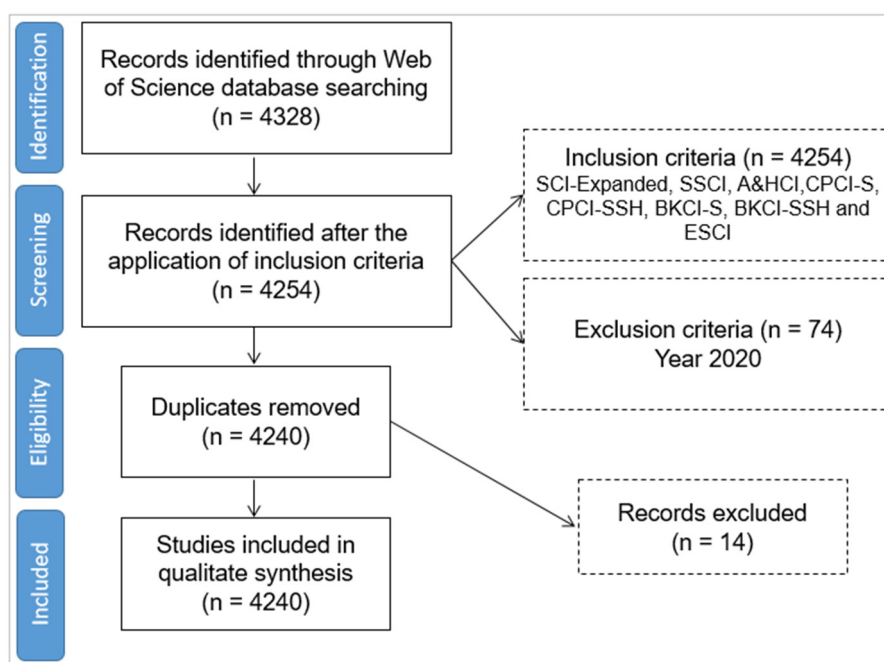
This study focused on the analysis of co-words [31] and bibliometric indicators such as the h-index and derivatives (g, hg, q2) [32]. This study allowed us to generate maps with nodes that determined the performance and location of various conceptual subdomains linked to the terms “Machine Learning” and “Big Data”. This served to specify the thematic development of these constructs in WoS [33].

### 3.2. Procedure and Data Analysis

The research process was carried out in different actions. First the database was selected. In this case, WoS was chosen as a database that contains a large number of indexed impact studies. Next, the keywords to be analyzed were determined. In this study the terms “Machine Learning” and “Big Data” were chosen, after consultation in various specialized thesauri. Next, the search equation was constructed. The result was “Machine Learning” [TOPIC] AND “Big Data” [TOPIC] with the intention of refining the process of reporting scientific documents that had such terms in title, abstract, and keywords of indexed publications.

These first actions obtained a scientific production of 4328 documents. The first studies dated back to the year 2010. Therefore, the literature of the last 10 years (2010-2019) was taken, suppressing studies published in 2020 (n = 74) for not having finished the year and duplicates or indexed incorrectly (n = 14). Therefore, the unit of analysis focused on 4240 documents. This figure was the result of the application of various production indicators with their respective inclusion criteria such as year of publication (all production except 2020), language (x ≥ 10), publication area (x ≥ 700), type of documents (x ≥ 100), organizations (x ≥ 50), authors (x ≥ 10), sources of origin (x ≥ 30), countries (x ≥ 200), citation (the four most cited documents; x ≥ 250). The monitoring of these actions resulted

in the generation of the following flow chart based on the protocols of the PRISMA-P (PRISMA for systematic review protocols (PRISMA-P) matrix (Figure 1).



**Figure 1.** Flowchart according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Declaration.

To analyze the reported literature, various software was used. Two are tools from WoS, Analyze Results and Creation Citation Report. These were used to extract the data related to the year, authorship, country, type of document, institution, language, medium and most cited documents. The other program was SciMAT, used to longitudinally analyze the structural and dynamic development of scientific production. For an effective analysis, the instructions of experts in this latest software were followed [34]. SciMAT allowed the following thematic co-word analysis to be carried out through the following processes:

**Recognition:** in this process various actions were carried out: a) analyze the keywords of the reported documents ( $n = 12657$ ); b) generate a map of co-occurrence nodes; c) Develop a standardized network of co-words; d) Detect the keywords with greater significance ( $n = 11993$ ); e) Represent the most influential topics and terms through a clustering algorithm.

**Reproduction:** Following the principles of centrality and density, a strategic diagram and a thematic network were developed. Centrality measures the degree of interaction of a network with other networks and is expressed by the equation  $c=10 \cdot \sum e_{kh}$ , where  $k$  is a keyword belonging to the topic and  $h$  a keyword belonging to other topics. Centrality analyses the strength of external links to other themes.

This value was considered as the measure of the importance of a theme in the development of the entire field of research analyzed. Density measures the internal strength of the network and is expressed by the equation  $d = 100 \cdot \sum e_{ij}/w$ , where  $i$  and  $j$  are keywords belonging to the topic and  $w$  is the number of keywords in the topic. Density analyzes the strength of the internal links between all the keywords that describe the research topic. This value was considered a measure of the degree of development of the topic under study. In the graphic study generated, there were four areas: upper right (motor and relevant topics), upper left (rooted and isolated themes), lower left (missing or projected topics) and lower right (low development and transversal themes).

**Determination:** The development of the nodes in different periods or time intervals is studied. In this case, five periods were delimited ( $P_1 = 2010-2015$ ;  $P_2 = 2016$ ;  $P_3 = 2017$ ;  $P_4 = 2018$ ;  $P_5 = 2019$ ).

The strength of association was achieved through the volume of keywords in common in the different periods. However, for the authorship all literary production was taken. Therefore, a single period was configured ( $P_X = 2010-2019$ ).

Performance: To carry out this process, various production indicators were taken with their corresponding inclusion criteria in order to be considered in the study (Table 1).

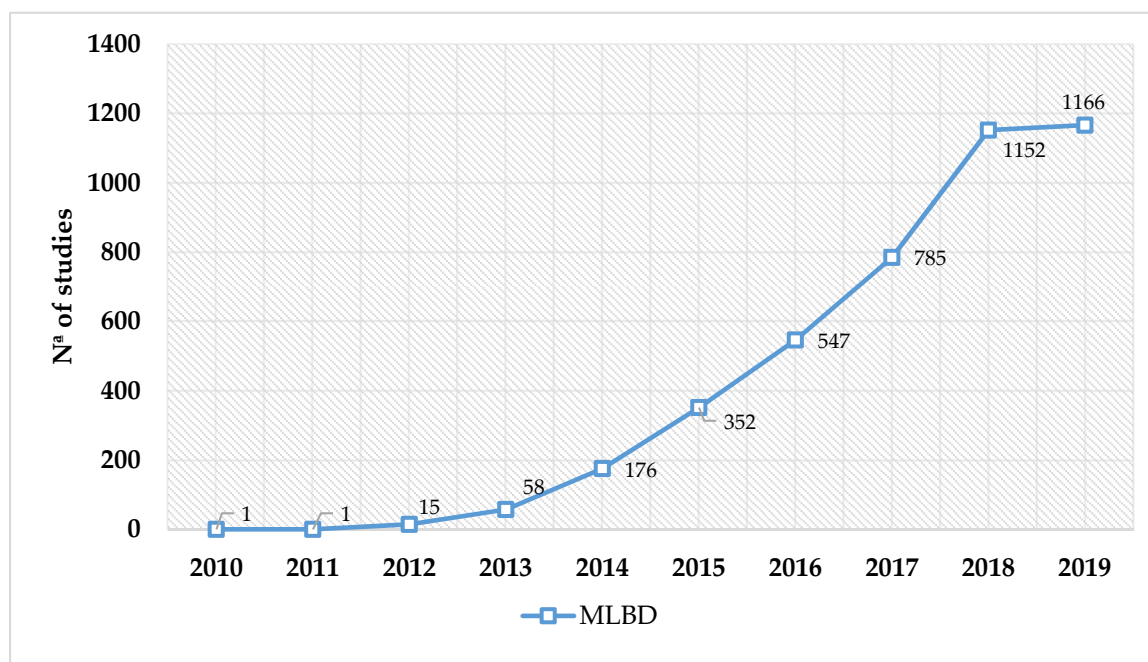
**Table 1.** Production indicators and inclusion criteria.

Configuration	Values
Analysis unit	Keywords authors, keywords WoS
Frequency threshold	Keywords: $P_1 = (3)$ , $P_2 = (2)$ , $P_3 = (3)$ , $P_4 = (6)$ , $P_5 = (6)$ Authors: $P_X = (4)$
Network type	Co-occurrence
Co-occurrence union value threshold	Keywords: $P_1 = (2)$ , $P_2 = (2)$ , $P_3 = (2)$ , $P_4 = (2)$ , $P_5 = (2)$ Authors: $P_X = (3)$
Normalization measure	Equivalence index
Clustering algorithm	Maximum size: 9; Minimum size: 3
Evolutionary measure	Jaccard index
Overlapping measure	Inclusion Rate

## 4. Results

### 4.1. Performance and Scientific Production

The evolution of the 4240 documents in the scientific production on MLBD has been constant and continuous in the time, having an exponential growth from its beginnings until the year 2018, where they maintained a stable level of production until the year 2019. In other words, the production levels in 2018 and 2019 were even, showing an equal interest in both years by the scientific community (Figure 2).



**Figure 2.** Evolution of scientific production of diet in education in the Web of Science (WoS).

The language chosen by authors for the presentation of the academic results was mostly English (Table 2a). The main areas of knowledge in MLBD studies were maintained with even numbers in

Computer Science Theory Methods, Computer Science Information Systems and Engineering Electrical Electronic (Table 2b).

**Table 2.** Descriptive bibliometric variables.

<b>Production by Language (a)</b>	<b>n</b>
English	4187
Turkish	16
German	14
Spanish	12
<b>Production by research area (b)</b>	<b>n</b>
Computer Science Theory Methods	1041
Computer Science Information Systems	1010
Engineering Electrical Electronic	1000
Computer Science Artificial Intelligence	738
<b>Production by document types (c)</b>	<b>n</b>
Article	1934
Proceedings paper	1856
Review	356
Editorial Material	114
<b>Production by institution (d)</b>	<b>n</b>
University of California System	122
Harvard University	77
Chinese Academy of Sciences	76
University of Texas System	66
State University System of Florida	63
<b>Production by authors (e)</b>	<b>n</b>
Wang L.	18
Wang Y.	18
Zhang Y.	15
Liu Y.	14
Lee S.	13
Zhang L.	13
Li X.	12
Liu H.	12
Kim Y.	11
Kumar A.	11
<b>Production by source (f)</b>	<b>n</b>
IEEE International Conference on Big Data	165
IEEE Access	94
Lecture Notes in Computer Science	77
Procedia Computer Science	36
<b>Production by countries (g)</b>	<b>n</b>
United States	1479
China	729
India	311
England	302
Germany	213

There were even numbers in the type of document used to present the information, being used mainly the articles and the communications in congresses (Table 2c). The main organization that referred to MLBD studies was the University of California Systems, being quite distant from the rest (Table 2d).



The authors with the highest production were Wang L. and Wang Y., there being no great differences with the rest of the authors (Table 2e). The main source of presentation of studies on MLBD was the IEEE International Conference on Big Data, which gathered the compilation of works developed in congresses.

The main journal was Lecture Notes in Computer Science, which was the main producer in this field of study (Table 2f). The country with the greatest interest in production over MLBD was the United States, with twice as much production as the next country, China (Table 2g).

The reference authors for the scientific community, due to his high citation, was Kosinski, M.; Stillwell, D.; Graepel, T., with their article titled “Private traits and attributes are predictable from digital records of human behaviour”, who accumulated a high number of citations in the MLBD study. These authors were followed, with fewer citations, by Muja, M.; Lowe, D.G., with their article titled “Scalable Nearest Neighbor Algorithms for High Dimensional Data”. (Table 3).

Table 3. Most cited articles.

Reference	Citations
Kosinski, M.; Stillwell, D.; Graepel, T. Private traits and attributes are predictable from digital records of human behaviour. <i>Proceedings of the national academy of sciences of the United States of America</i> <b>2013</b> , <i>110</i> , 5802-5805. doi: 10.1073/pnas.1218772110	607
Muja, M.; Lowe, D.G. Scalable Nearest Neighbor Algorithms for High Dimensional Data. <i>IEEE Transactions on Pattern Analysis and Machine Intelligence</i> <b>2014</b> , <i>36</i> , 2227-2240. doi: 10.1109/TPAMI.2014.2321376	455
Obermeyer, Z.; Emanuel, E.J. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. <i>New England Journal of Medicine</i> <b>2016</b> , <i>375</i> , 1216-1219. doi: 10.1056/NEJMp1606181	397
Chen, X.W.; Lin, X. Big Data Deep Learning: Challenges and Perspectives. <i>IEEE Access</i> <b>2014</b> , <i>2</i> , 514-525. Doi: 10.1109/ACCESS.2014.2325029	272

#### 4.2. Structural and Thematic Development

The evolution of keywords shows information about the number of keywords in each of the established time intervals, the number of matching keywords between the periods and the number of keywords leaving and entering a certain period with respect to another. In this case, a more established line of research can be observed in the last four years, which shows the same trend in the scientific community itself (Figure 3).

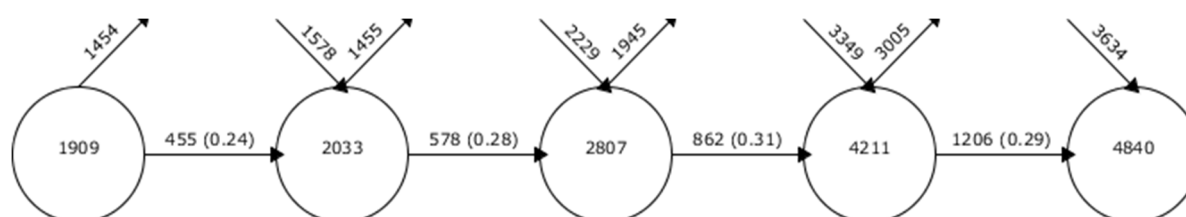


Figure 3. Continuity of keywords among contiguous intervals.

The academic performance in the established periods offers the subjects with the greatest bibliometric indicators, using the h index as the main reference, and completing this information with the g index, hg index and  $q^2$  index, in addition to the number of citations.

In this case, the “machine-learning” theme was shown to be the one that presents the highest bibliometric indicators in all periods, except in 2016, where “predictions” was the theme with the highest values. The variety of themes that appeared in the different periods is noteworthy, offering the main lines of research developed (Table 4).

Table 4. Thematic performance.

Interval 2010–2015						
Denomination	Works	Index-h	Index-g	Index-hg	Index-q <sup>2</sup>	Citations
Machine-learning	165	22	46	31.81	31.81	2463
Algorithm	14	9	14	11.22	13.42	318
Prediction	17	7	15	10.25	10.91	272
Big-Data_Analytics	18	7	13	9.54	15.2	300
Neural-Networks	7	5	6	5.48	15	446
Hadoop	8	5	5	5	15.49	229
Data_streams	5	3	4	3.46	3.46	20
Sparse-Representation	4	3	4	3.46	10.68	111
Sentiment-Analitics	6	2	4	2.83	8.72	45
Support-Vector-Machine	4	2	4	2.83	6.32	30
Interval 2016						
Denomination	Works	Index-h	Index-g	Index-hg	Index-q <sup>2</sup>	Citations
Predictions	31	16	27	20.78	21.54	766
Data-mining	39	11	26	16.91	17.23	706
Classification	22	9	19	13.08	17.75	397
Networks	12	9	12	10.39	14.07	533
Neural-networks	13	6	13	8.83	12.49	320
Model	8	6	7	6.48	15.49	212
Natural-language-processing	9	4	7	5.29	11.31	291
Mapreduce	12	4	6	4.9	6.63	46
Dynamics	4	4	4	4	12	97
Analytics	5	4	5	4.47	9.8	105
Mass-spectrometry	3	3	3	3	9.8	88
Smart-Meter	3	3	3	3	3	46
Language	4	1	3	1.73	5.66	34
Lung-Cancer	2	2	2	2	11.92	88
Privacy	4	2	2	2	16.37	241
ITS	3	1	1	1	2.45	6
Harnessing-interference	2	1	1	1	3.46	12
Interval 2017						
Denomination	Works	Index-h	Index-g	Index-hg	Index-q <sup>2</sup>	Citations
Machine-learning	155	22	43	30.76	31.11	2022
System	16	9	16	12	13.42	377
Algorithm	29	9	22	14.07	15.87	514
Supporter-Vector-Machine	15	8	13	10.2	13.56	434
Social-Media	16	8	12	9.8	10.58	167
Framework	12	6	9	7.35	7.75	342
Mapreduce	43	6	8	6.93	8.12	104
Surveillance	5	3	3	3	4.24	21
Features	4	3	3	3	6.48	239
Cloud	6	3	4	3.46	8.31	58
Apache-Spark	13	2	5	3.16	7.35	42
Analytics	5	2	4	2.83	14.83	119
Prevention	3	1	1	1	1.73	3
Text-mining	4	1	2	1.41	3.46	13

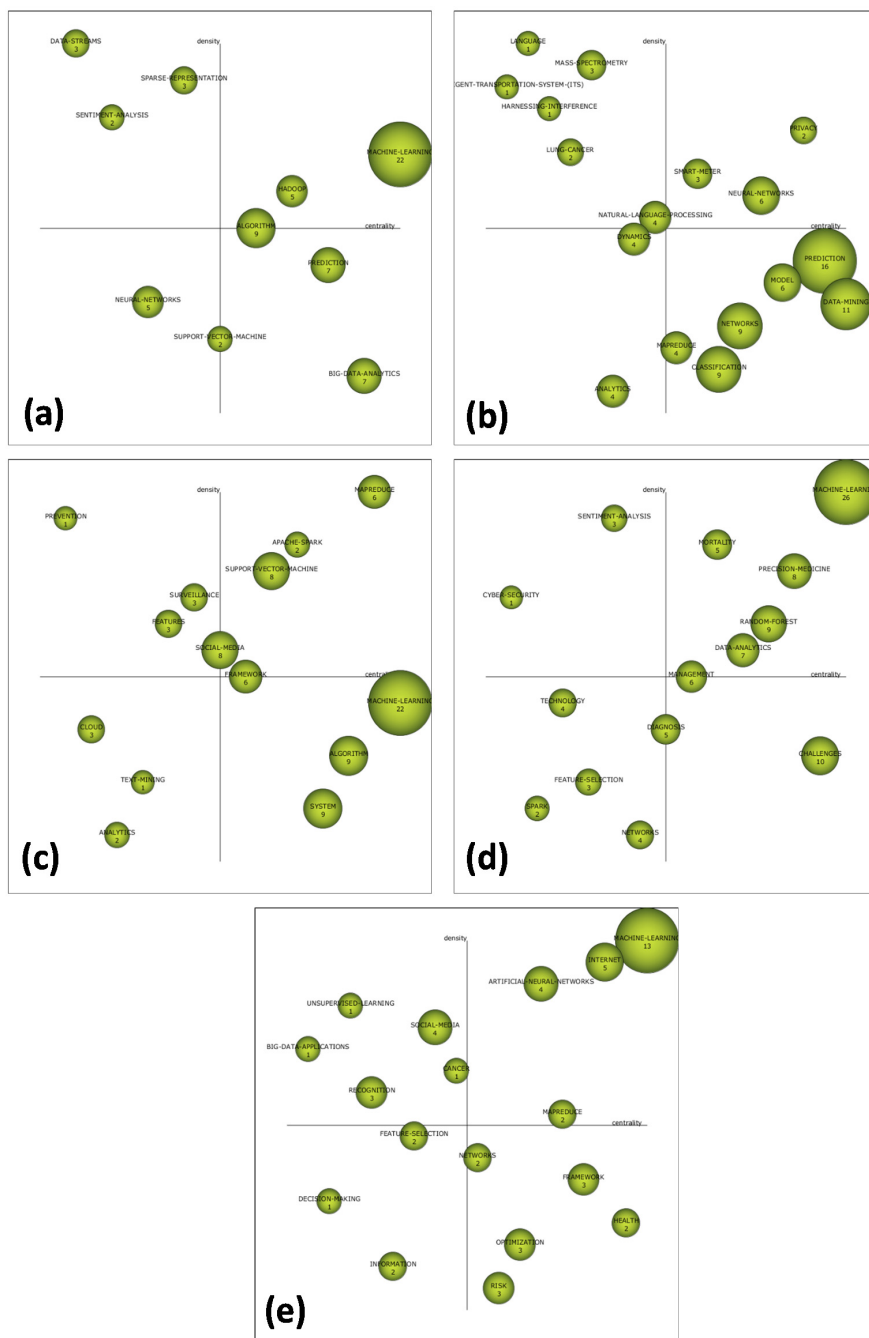


Table 4. Cont.

Interval 2018						
Denomination	Works	Index-h	Index-g	Index-hg	Index-q <sup>2</sup>	Citations
Machine-learning	484	26	39	31.84	31.84	2895
Callenges	22	10	17	13.04	17.32	378
Random-Forest	28	9	19	13.08	15	385
Precision-Medicine	23	8	14	10.58	12.33	213
Data-analytics	22	7	16	10.58	12.69	267
Managements	16	6	11	8.12	11.75	142
Mortaly	11	5	8	6.32	8.37	74
Diagnosis	14	5	9	6.71	7.42	99
Networks	8	4	5	4.47	10.39	83
Technology	5	4	4	4	6	39
Sentiment-analysis	24	3	7	4.58	6.24	63
Feature-Selection	8	3	6	4.24	5.48	47
Spark	6	2	3	2.45	5.1	19
Cyber-Security	3	1	2	1.41	4.8	24
Interval 2019						
Denomination	Works	Index-h	Index-g	Index-hg	Index-q <sup>2</sup>	Citations
Machine-learning	519	13	19	15.72	16.52	948
Internet	37	5	8	6.32	5	86
Artificial-neural-networks	19	4	6	4.9	6.93	54
Social-media	20	4	7	5.29	5.29	60
Framework	21	3	3	3	3.46	26
Optimization	23	3	5	3.87	3.87	35
Risk	12	3	7	4.58	7.94	88
Recognition	11	3	4	3.46	3	19
Feature-selection	13	2	2	2	2.83	10
Networks	15	2	2	2	2.83	7
Mapreduce	31	2	5	3.16	6.48	36
Health	11	2	3	2.45	3.16	15
Information	9	2	3	2.45	3.74	10
Cancer	13	1	2	1.41	2	8
Unsupervised-learning	6	1	2	1.41	2	8
Big-Data-Applications	5	1	2	1.41	2	6
Decision-Making	4	1	1	1	4.9	24

The diagrams of the intervals developed show data on the importance of each of the themes in the different periods. For this purpose, a grouping process was developed, according to Callon's indicators, which assesses the degree of interaction of a network with respect to other networks, from two axes: centrality, which analyzes the strength of the relationship of external links with other topics, where it shows the importance of the development of a topic in a field of research; and density, which assesses the internal strength of the network, analysing the internal links between the key words that are grouped around a specific topic, giving information on the degree of development of a field of study. In the first period (2010-2015), the driving themes were "machine-learning" and "Hadoop".

In the second period (2016), the driving themes were "privacy", "Smart-meter" and "neural-networks". In the third period (2017) it was "mapreduce", "apache-spark" and "support-vector-machine". In the fourth period, it was "machine-learning", "mortality", "precision-medicine", "random-forest" and "data-analytics". In the last period (2019) it was "machine-learning", "internet", "artificial-neural-networks" and "mapreduce". In this period, we must bear in mind the themes "feature-selection", "decision-making" and "information", given that their location in the diagram makes them unknown, given that they may be the driving force in the future or may tend to disappear from scientific production (Figure 4).



**Figure 4.** Machine Learning Big Data (MLBD)'s h-indexed strategic diagram. Note: (a) Interval 2010-2015; (b) interval 2016; (c) interval 2017; (d) interval 2018; (e) interval 2019.

### 4.3. Thematic Evolution of the Terms

The thematic evolution analyzes the thematic development of the scientific field studied, according to the number of established time periods. In this case,  $T_t$  is the set of themes detected for a given period, where  $U \in T_t$  represents each of the themes detected in period  $X$ . Let  $V \in T_{t+1}$  be the set of themes detected in the following period of time  $x+1$ . In this case, it can be determined that there was thematic evolution from theme  $U$  to theme  $V$  if there were thematic networks of both themes, in which at least one keyword was shared. In this way,  $V$  could be considered an evolved theme from  $U$ . The keywords  $k \in U \cap V$  were considered as thematic nexus or conceptual nexus of evolution.

The importance of thematic nexus was measured by the number of themes they had in common, measured in solid lines and dotted lines. Solid lines mean that the linked topics shared the same name,

that is, both topics are labelled with the same keyword, or the label of one of the topics was part of the other topic. A dashed line means that the topics shared elements other than the name of the topics. The strength of the links between two topics is proportional to the value of the Jaccard index of both topics. The volume of the spheres is proportional to the number of documents associated with the topic.

The results show that a conceptual gap existed if all periods were taken into account, given that there were no themes that were repeated in all established intervals. In this case, the year 2016 was the one that produces this gap, given that the rest of the periods the conceptual line marks “machine-learning”, especially in the last three years, where the connection, besides being thematic, was solid and consolidated, placing it as a reference in this field of research.

A relevant aspect to bear in mind is that there are more thematic connections than key words, which shows that the trends in research are established and connected. The evolution of the studies on MLBD determines how the studies, in the first years were based on purely computer science aspects, evolving towards medical aspects, thus showing the transformation of the research in this field of study (Figure 5).

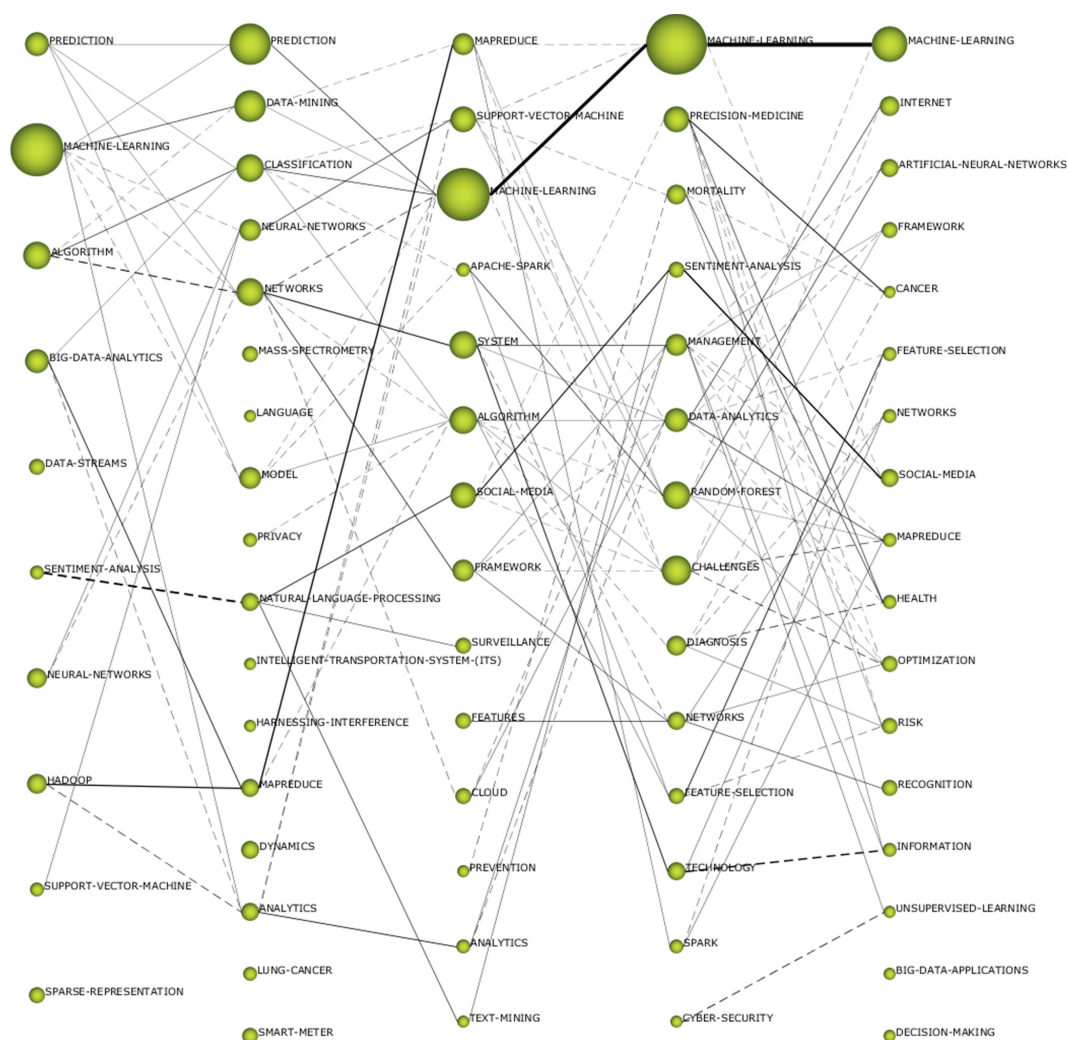


Figure 5. Thematic evolution by h-index.

#### 4.4. Authors with a Higher Relevance Index

According to the data shown in the authors study, Peixoto, R. or Poornachandran, P. were positioned as the driving authors, while Song, J.N. was positioned as an author who may be relevant in the future in this field of study. There were authors such as Passos, I.C., Momayoun, H. and Mosavi,

A., who although they showed the highest h indexes, were both developed and isolated, or were basic and transversal (Figure 6).

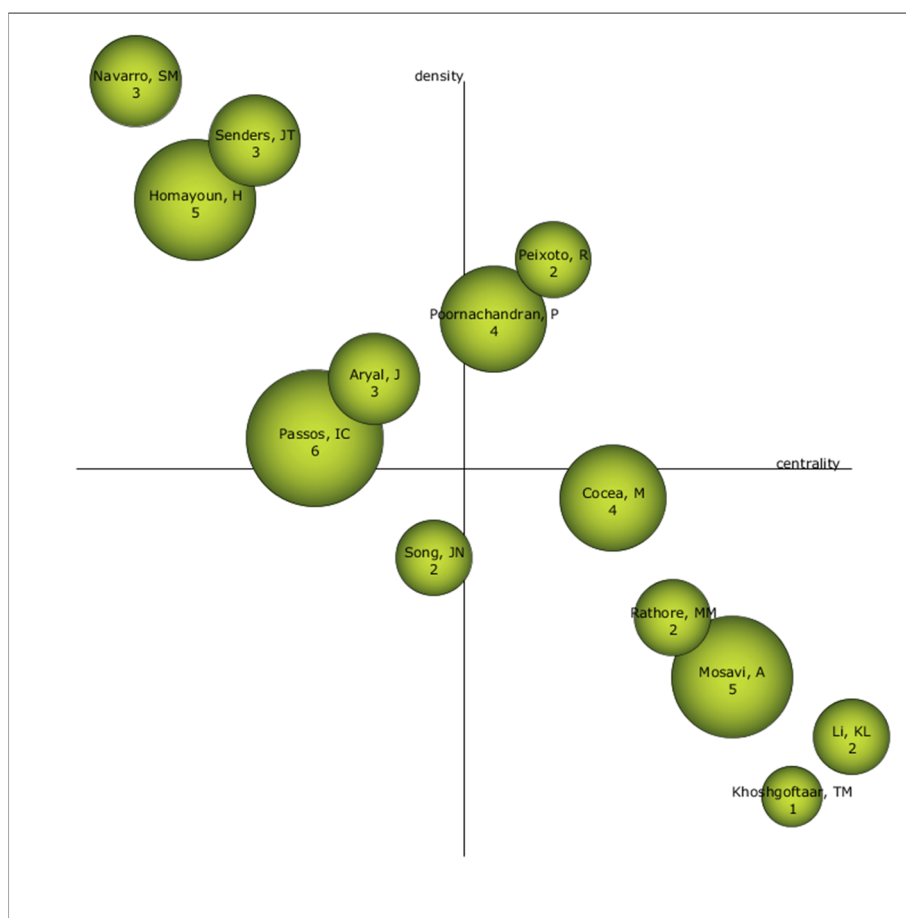


Figure 6. Strategic authoring diagram.

## 5. Discussion

As it has been shown in the previous section, there was an exponential growth within MLBD publications from 2010 to 2018, thus maintaining the number of publications the following year [35]. This reflects a state-of-the-art that is booming and is of interest to researchers from the scientific community who are contributing to the advancement of this field of knowledge and, in the same way, to science.

In a firm way, as it happens in other studies [36–38], the language of publications is mostly English, as a worldwide scientific language, and publications with very striking figures in other languages are relegated, not reaching significant and outstanding values such as Anglo-Saxon, which is situated as the predominant language.

Most of the published documents are articles and communications in congresses, in order to be able to disseminate the studies on MLBD [9]. These ones reach an outstanding figure, hovering over almost two thousand documents each, while the rest of the type of documents such as material review and editorial obtain minority figures in comparison.

The most productive organization on MLBD studies is the University of California Systems, with a great difference from other organizations. As in other studies on other topics [39]. The United States remains the richest country in MLBD productions, followed by China, with almost half of its number of publications. Regarding the study that has received the most citations, it is called Private traits and attributes are predictable from digital records of human behavior, published in Proceedings of the national academy of sciences of the United States of America in 2013, by the authors

Kosinski, M., Stillwell, D., and Graepel, T. [35], exceeding half a thousand citations. On the other hand, it is determined that the machine-learning theme has the greatest bibliometric indicators in the periods analyzed.

## 6. Conclusions

After the analysis, it is shown that there are more connections between the themes than between the keywords themselves, so important for discourse analysis, which reveals that the research trend is related to studies. The publications on the subject of this research have been more prolific during the last four years, which is where there is a greater coincidence of keywords between periods. When considering a theme par excellence in this type of studies, it is “machine-learning”, which has the highest bibliometric levels in most of the analyzed times and is the one that most appears as a motor theme in the established periods. This shows the great importance of the term on the part of the scientific community when carrying out its investigations, being even above Big Data.

Research in this field also shows that there are many thematic connections between them, being able to elucidate that the studies are related to each other. In addition, it shows an evolution in the base of the studies, going at first based on purely medical aspects, advancing in recent times to aspects related to the field of medicine. Finally, the authors Peixoto, R. or Poornachandran, P. are placed as motor authors, and therefore, those that have more relevance and importance for the educational community.

As a future prospect, the idea arises to propose an in-depth analysis of content on MLBD publications, analyzing whether the trend in the texts arises as investigations with or without experiences, and/or if they are only at a theoretical level and if these publications maintain the defence in a positive or negative direction of the subject analyzed.

There are several limitations presented in this investigation. First, there is the debugging of the data presented in the WoS, where repeated documents are presented or that are not related to the subject of the study. Second, the establishment of the intervals, in this case a matter of fairness, since researchers have always tried to maintain a similar number of documents in each of the intervals. Third and last, the parameters marked in this study have been established according to the researchers' own criteria, who have tried to present the results according to their size and relevance. Therefore, the data presented here should be analyzed with caution, since changing the parameters established in this investigation may lead to a variation in the number and connections in the subjects presented.

**Author Contributions:** Conceptualization, A.S.-R.; methodology, J.L.B.; software, A.-J.M.-G.; formal analysis, J.L.B., A.-J.M.-G., M.E.P.-G. and A.S.-R.; investigation, J.L.B., A.-J.M.-G., M.E.P.-G. and A.S.-R.; data curation, A.-J.M.-G.; writing—original draft preparation, J.L.B., A.-J.M.-G., M.E.P.-G. and A.S.-R.; writing—review and editing, J.L.B., A.-J.M.-G., M.E.P.-G. and A.S.-R.; visualization, A.-J.M.-G., M.E.P.-G. and A.S.-R.; supervision, J.L.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** The author received no specific funding for this research.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Qiu, J.; Wu, Q.; Ding, G.; Xu, Y.; Feng, S. Erratum to: A survey of machine learning for big data processing. *EURASIP J. Adv. Signal Process.* **2016**, *1*, 1–16. [[CrossRef](#)]
2. Zhou, L.; Pan, S.; Wang, J.; Vasilakos, A.V. Machine learning on big data: Opportunities and challenges. *Neurocomputing* **2017**, *237*, 350–361. [[CrossRef](#)]
3. Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials discovery and design using machine learning. *J. Mater.* **2017**, *3*, 159–177. [[CrossRef](#)]
4. Das, S.; Dey, A.; Pal, A.; Roy, N. Applications of Artificial Intelligence in Machine Learning: Review and Prospect. *IJCA* **2015**, *115*, 31–41. [[CrossRef](#)]
5. Fan, W.; Bifet, A. Mining big data: Current status, and forecast to the future. *SIGKDD Explor. Newsl.* **2013**, *14*, 1–5. [[CrossRef](#)]

6. Fan, S.K.S.; Su, C.J.; Nien, H.T.; Tsai, P.F.; Cheng, C.Y. Using machine learning and big data approaches to predict travel time based on historical and real-time data from Taiwan electronic toll collection. *Soft. Comput.* **2018**, *22*, 5707–5718. [[CrossRef](#)]
7. Hanzelik, P.P.; Gergely, S.; Gáspár, C.; Győry, L. Machine learning methods to predict solubilities of rock samples. *J. Chemom.* **2020**, *34*, 1–13. [[CrossRef](#)]
8. Jena, R.K. Sentiment mining in a collaborative learning environment: Capitalising on big data. *Behav. Inf. Technol.* **2019**, *38*, 986–1001. [[CrossRef](#)]
9. Boyd, D.; Crawford, K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Inf. Commun. Soc.* **2012**, *15*, 662–679. [[CrossRef](#)]
10. Daelemans, W.; Hoste, V. Evaluation of machine learning methods for natural language processing tasks. In Proceedings of the LREC 2002 Third international conference on language resources and evaluation; European Language Resources Association (ELRA), Las Palmas de Gran Canaria, Spain, 29–31 May 2002; p. 6.
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)]
12. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [[CrossRef](#)] [[PubMed](#)]
13. Menshaw, A. *Deep Learning by Example: A Hands-on Guide to Implementing Advanced Machine Learning Algorithms and Neural Networks*, 1st ed.; Packt Publishing: Birmingham, UK, 2018.
14. Bhardwaj, A.; Di, W.; Wei, J. *Deep Learning Essentials: Your Hands-on Guide to the Fundamentals of Deep Learning and Neural Network Modeling*, 1st ed.; Packt Publishing: Birmingham, UK, 2018.
15. Alaei, A.R.; Becken, S.; Stantic, B. Sentiment Analysis in Tourism: Capitalizing on Big Data. *J. Travel Res.* **2019**, *58*, 175–191. [[CrossRef](#)]
16. Kraus, M.; Feuerriegel, S.; Oztekin, A. Deep learning in business analytics and operations research: Models, applications and managerial implications. *Eur. J. Oper. Res.* **2020**, *281*, 628–641. [[CrossRef](#)]
17. Zhang, W.; Wang, P.; Sun, K.; Wang, C.; Diao, D. Intelligently detecting and identifying liquids leakage combining triboelectric nanogenerator based self-powered sensor with machine learning. *Nano Energy* **2019**, *56*, 277–285. [[CrossRef](#)]
18. Zhang, Q.; Yang, L.T.; Chen, Z.; Li, P. A survey on deep learning for big data. *Inf. Fusion* **2018**, *42*, 146–157. [[CrossRef](#)]
19. Serrano, E.; Bajo, J. Deep neural network architectures for social services diagnosis in smart cities. *Future Gener. Comput. Syst.* **2019**, *100*, 122–131. [[CrossRef](#)]
20. Gök, A.; Waterworth, A.; Shapira, P. Use of web mining in studying innovation. *Scientometrics* **2015**, *102*, 653–671. [[CrossRef](#)]
21. Montáns, F.J.; Chinesta, F.; Gómez-Bombarelli, R.; Kutz, J.N. Data-driven modeling and learning in science and engineering. *Comptes Rendus Mécanique* **2019**, *347*, 845–855. [[CrossRef](#)]
22. Liang, X.; Fan, L.; Loh, Y.P.; Liu, Y.; Tong, S. Happy Travelers Take Big Pictures: A Psychological Study with Machine Learning and Big Data. *arXiv* **2017**, arXiv:1709.07584.
23. Manogaran, G.; Vijayakumar, V.; Varatharajan, R.; Malarvizhi, P.; Sundarasekar, R.; Hsu, C.H. Machine Learning Based Big Data Processing Framework for Cancer Diagnosis Using Hidden Markov Model and GM Clustering. *Wirel. Pers. Commun.* **2018**, *102*, 2099–2116. [[CrossRef](#)]
24. Jan, B.; Farman, H.; Khan, M.; Imran, M.; Islam, I.U.; Ahmad, A.; Ali, S.; Jeon, G. Deep learning in big data Analytics: A comparative study. *Comput. Electr. Eng.* **2019**, *75*, 275–287. [[CrossRef](#)]
25. Gómez-García, A.; Ramiro, M.T.; Ariza, T.; Granados, M.R. Estudio bibliométrico de Educación XX1. *Educ. XX1* **2012**, *15*, 17–41.
26. Montilla, L.J. Análisis bibliométrico sobre la producción científica archivística en la Red de Revistas Científicas de América Latina y el Caribe (Redalyc) durante el período 2001–2011. *Biblios* **2012**, *48*, 1–11. [[CrossRef](#)]
27. López-Belmonte, J.; Moreno-Guerrero, A.J.; López-Núñez, J.A.; Pozo-Sánchez, S. Analysis of the Productive, Structural, and Dynamic Development of Augmented Reality in Higher Education Research on the Web of Science. *Appl. Sci.* **2019**, *9*, 5306. [[CrossRef](#)]
28. Rodríguez-García, A.-M.; López Belmonte, J.; Agreda Montoro, M.; Moreno-Guerrero, A.J. Productive, Structural and Dynamic Study of the Concept of Sustainability in the Educational Field. *Sustainability* **2019**, *11*, 5613. [[CrossRef](#)]



29. Martínez, M.A.; Cobo, M.J.; Herrera, M.; Herrera-Viedma, E. Analyzing the Scientific Evolution of Social Work Using Science Mapping. *Res. Soc. Work Pract.* **2015**, *25*, 257–277. [[CrossRef](#)]
30. Moral-Muñoz, J.A.; Herrera-Viedma, E.; Santisteban-Espejo, A.; Cobo, M.J. Software tools for conducting bibliometric analysis in science: An up-to-date review. *EPI* **2020**, *29*, 1–20. [[CrossRef](#)]
31. Hirsch, J.E. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16569–16572. [[CrossRef](#)] [[PubMed](#)]
32. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. Science mapping software tools: Review, analysis, and cooperative study among tools. *J. Am. Soc. Inf. Sci.* **2011**, *62*, 1382–1402. [[CrossRef](#)]
33. López-Robles, J.R.; Otegi-Olaso, J.R.; Porto Gómez, I.; Cobo, M.J. 30 years of intelligence models in management and business: A bibliometric review. *Int. J. Inf. Manag.* **2019**, *48*, 22–38. [[CrossRef](#)]
34. Montero-Díaz, J.; Cobo, M.J.; Gutiérrez-Salcedo, M.; Segado-Boj, F.; Herrera-Viedma, E. A science mapping analysis of 'Communication' WoS subject category (1980–2013). *Comun. Rev. Científica Comun. Educ.* **2018**, *26*, 81–91.
35. Kosinski, M.; Stillwell, D.; Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5802–5805. [[CrossRef](#)] [[PubMed](#)]
36. Torres, K. Tendencia en la Transformación Digital Para Retailers: Omnicanalidad Soportada Por "Big Data Analytics" Para Mejorar la Experiencia del Cliente Durante su Recorrido: Análisis de Adopción en Argentina. Ph.D. Thesis, Universidad de San Andrés, Victoria, Argentina, 2017.
37. Parra-González, M.E.; Segura-Robles, A. Producción científica sobre gamificación en educación: Un análisis cuantitativo. *Rev. Educ.* **2019**, *5*, 113–131.
38. Aguado-López, G.; Rogel-Salazar, E.; Becerril-García, R.; Baca-Zapata, A. Presencia de universidades en la red: La brecha digital entre Estados Unidos y el resto del mundo. *RUSC Univ. Knowl. Soc. J.* **2009**, *6*, 1–18.
39. Rodríguez, Á.; Mas, L. Inventario de palabras clave temáticas para la clasificación automática de noticias de televisión. *An. Doc.* **2011**, *14*, 1–24.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).