mathematics

Article

Data Analysis and Domain Knowledge for Strategic Competencies Using Business Intelligence and Analytics

Mauricio Olivares Faúndez [1,*] and Hanns de la Fuente-Mella [2,*]

[1] Facultad de Economía y Negocios, Universidad Finis Terrae, Santiago 7501015, Chile
[2] Instituto de Estadística, Facultad de Ciencias, Pontificia Universidad Católica de Valparaíso, Valparaíso 2340031, Chile
* Correspondence: molivares@uft.cl (M.O.F.); hanns.delafuente@pucv.cl (H.d.l.F.-M.)

Abstract: This research arises from the demand in business management for capabilities that put into practice—in an autonomous way—skills and knowledge in BI&A of all those who make decisions and lead organizations. To this end, this study aims to analyze the development of scientific production over the last 20 years in order to provide evidence of possible gaps, patterns and emphasis on domains of strategic leadership competencies in BI&A. The study was split into two methodological phases. Methodological Phase 1: Application of analytical techniques of informetrics. Methodological Phase 2: natural language processing and machine learning techniques. The records collected were 1231 articles from the Web of Science and Scopus databases on 16 August 2021. The results confirm, with an r^2 = 96.9%, that a small group of authors published the largest number of articles on strategic leadership competencies in BI&A. There is also a strong emphasis on studies in the domain of professional capability development (92.29%), and there are few studies in the domain of enabling environment for learning (0.72%); the domain of expertise (3.01%) and strategic vision of BI&A was also rare (3.37%).

Keywords: informetric; intelligence and analytics; competencies; capacities; machine learning (ML); formalization of domain knowledge

MSC: 68T05

check for updates

Citation: Faúndez, M.O.; de la Fuente-Mella, H. Data Analysis and Domain Knowledge for Strategic Competencies Using Business Intelligence and Analytics. *Mathematics* **2023**, *11*, 34. https://doi.org/10.3390/math11010034

Academic Editor: Jan Rauch

Received: 21 November 2022
Revised: 15 December 2022
Accepted: 17 December 2022
Published: 22 December 2022

1. Introduction

Business intelligence and analytics (BI&A) is an umbrella term commonly used to describe the technologies, applications, and processes for collecting, storing, accessing, and analyzing data to help users make better decisions, and best practices for information analysis [1,2]. BI&A has shifted from producing static reports to generating real-time, integrated information. Conceptually, BI&A is divided into three phases. BI&A 1.0 was centered around descriptive analytics, where data were structured and collected from within companies. In BI&A 2.0, big data appeared, making BI&A a new strategic phase for understanding market needs [3]. With the increasing development of IT—the rise in web and mobile devices—BI&A experienced explosive development, leading to BI&A 3.0, which presents the challenge of working with unstructured data [3,4]. The hallmark of BI&A 3.0 analytics is the extensive use of analytics by traditional enterprises leading to the potential for transformation of their business models and culture. In this stage, companies create large-scale data and analytics-based products, and analytics activities are becoming increasingly industrialized, often having thousands of machine learning (ML) models [5]. This has led to the rapid development of artificial intelligence (AI), which in recent years is gaining more attention because of the enormous amount of data available [6]. AI has tools and techniques for smarter, faster, and more actionable predictive analytics [7]. It has the ability to handle large amounts of data in real time, offering the possibility of results with great accuracy. Given that AI can enhance the value of BI&A, Davenport [5] argues that it should be considered an extension of a company's analytics capabilities. He claims it is leading the way to a new generation, BI&A 4.0, shifting the focus from descriptive to prescriptive and predictive analytics.

This technological and analytical development throughout the organization makes way for a more analytical culture, transforming and adapting organizational processes to manage the right information for the right people at the right time and fostering the skill to make less risky decisions in the organization. It is thus becoming clearer that the domain knowledge, that is, the knowledge of the field that the data belongs to, must be considered when analyzing data.

As with any major transition, executive leadership at the strategic level is necessary. In fact, the most important elements that decide the success or failure of BI&A in organizations include the quality of the data; the correct choice and implementation of technologies used; and development of analytical skills for human capital, sponsorship, and consistent alignment between BI and the strategic focus of the business, and its use [8,9]. In fact, mobilizing different capabilities of human capital for the development of leadership competencies can be learned, developed, and trained in daily experience [10–12]. According to organizational leadership reports, the two main strengths of the competencies approach are the flexibility and uniqueness of the concept to adapt to organizational needs [13,14]. Thus, the competencies necessary for successful leadership include a vision of the future, goal setting, communication, value fostering, the ability to gain insights into emerging visions, planning, and implementing a vision. These skills, along with managerial traits such as self-awareness, openness, self-confidence, and creativity serve as the basis for new leadership [15]. For effective leadership with high performance of organizations, it is important to develop cognitive and social intelligence competencies and emotional and behavioral skills of managers at all levels of the organizational structure, characterized by self-awareness, openness, self-confidence, and creativity in complex situations [16–18]. Due to the evolution of the global business environment in a context with a strong component of agility and high uncertainty, 21st-century organizations require leaders with the aforementioned competencies and skills incorporated into organizational management [19–21].

Our contribution aims to address four key aspects: firstly, we develop an analysis of the scientific production of the last 20 years on the competencies required by professionals who lead organizations by managing with BI&A, given that there are no studies that report on this. Secondly, the methodological design involved the use of two techniques for the development of the analysis, since all the studies in the field of BI&A, until now, have been limited to the use of bibliometrics, scientometrics, and informetrics, but none of them have corroborated their results with the use of unsupervised AI algorithm techniques. Thirdly, we report in what proportion the type of competencies to manage with BI&A is represented based on the analysis of key terms from the sweep of scientific output. The proportion of key terms is reported in relation to the dimension of competencies oriented to technical knowledge in BI&A, the dimension of competencies associated with organizational environment conducive to learning about BI&A, the dimension of competencies associated with Integrating BI&A skills into their own expert work (habits of mind) and the dimension of proportion of the scope of competencies that make it possible to achieve strategic business vision by creating value through BI&A. Fourthly, it shows the imbalance with respect to the studies in the scientific production of these 20 years in the dimensions of strategic competencies for leadership with BI&A, very relevant information to understand how to manage with BI&A in organizations, considering the great current and future development [22].

The paper is organized as follows: in Section 2, a review of the literature and previous work is developed; in Section 3, the problem is described; in Section 4, the study hypothesis is formulated and the methodology is developed; in Section 5, the results and main findings are presented; in Section 6, a discussion of the scope of the research is developed; and in Section 7, the limitations of this work are described and some directions for future research are outlined.

2. Literature Review and Previous Work

From an exhaustive review of the literature of the last 20 years, it is possible to note that there are no studies associated with the central theme of this research on strategic leadership competencies in BI&A. Only a few studies come close to this field. In that line, it is possible to mention the work of Wang [23] then in a review of the hermeneutic literature related to BI&A education, he stated that less attention has been paid to understanding the skill sets of various BI&A professionals and to demonstrating BI&A learning activity. In turn, Ardito, L. (2018) formed a bibliometric analysis of big data for business and management studied management and decision making in organizations [24]. Additionally, Di Vaio, A. (2022) developed a bibliometric analysis of humans' and artificial intelligence's effectiveness for public sector decision making [25]. Peifer, Jeske, and Hille (2022), in a study on artificial intelligence and its impact on leaders and leadership, pointed out that more research on its impact on leaders and leadership is needed to support companies with practice-tested guidelines and recommendations [26]. Thomas, et al. (2022), questioned whether the emotional intelligence competencies of management graduates predict their job performance, proving that several competencies of management graduates can predict the most important skill that can enable better job performance and reduce the employability gap, and in which a multi-layer artificial neural network tool was used to test the range of competencies among EQs, leadership quality, and work experience for perceived job performance [27]. Olszak (2022) pointed out that modern research indicates that the greatest impact on organizational development will pertain to business intelligence (BI). Indeed, it is believed that BI systems have become a strategic tool for economic growth, determining the competitiveness of many organizations and their innovative development. However, there is still very little research focused on exploring the problem of using BI systems in organizations [28].

This study in methodological stage 1 is based on the approach of analysis of the scientific production of the last 20 years on leadership competencies to perform with BI&A.

The term "informetrics" had its beginnings in the field of information science in the 1980s. In 1987, the International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval was held in Belgium. It suggested the inclusion of this term for the next conference to be held in London in 1989. The introduction of the word "informetrics" is attributed to the German Otto Nacke, who first used it in 1979 [29] At first it was only known as a general field of study that included elements of the earlier bibliometrics and scientometrics. Informetrics builds on the research of bibliometrics and scientometrics, and encompasses issues such as the development of theoretical models and measures of information, to find regularities in the data associated with the production and use of recorded information; it encompasses the measurement of aspects of information, storage and retrieval, and therefore, includes mathematical theory and modulation. Indeed, a descriptive study with informetric methodology examines the scientific production of competencies and professional skills in the field of BI&A knowledge in organizations. To achieve this purpose, we analyzed the evolution of scientific production, the productivity of authors according to Lotka's Law, the most productive journals, the analysis of the co-authorship map, and the co-occurrence of keywords to identify thematic trends. Lotka's law is a statistical observation offered by Alfred Lotka (1926) that describes a quantitative relationship between authors and articles produced in a given field and time period. In this sense, when Lotka's law is applied in this study, it shows that there is an uneven distribution, since most of the articles are written by a small portion of highly productive authors. Indeed, the number of authors A_n who publish papers on a subject is inversely proportional to the square of n: $A_n = \frac{A_1}{n^2}$, where A_n represents the number of papers by a given number of authors, A_1 is the number of papers produced by a single author and n^2 is the number of authors of the papers, so we apply the law of exponential growth squared [30].

The second methodological stage focused on methodologically applying two deep learning techniques using a Python library for natural language processing, and therefore, applying deep learning techniques for the extraction of textual aspects of scientific development capable of recognizing and generating main themes with accurate prediction—in this case, a technique called latent Dirichlet allocation (LDA). Effectively, the topic modeling involves extracting the characteristics of terms contained in research documents. Mathematical structures and context such as matrix factorization and singular value decomposition (SVD) are used, which leads to discovering some of the same type of information as the decomposition itself. This is achieved by generating a group or groups of terms that are differentiated from each other; groups of words form themes or concepts. Methodologically, these concepts are used to interpret the main themes of a corpus, to establish the semantic connections between words that appear frequently in many documents. There are several frameworks and algorithms for building topic models. One of these is the statistical technique of latent semantic indexing (LSI) for correlating semantically linked corpus terms. It is based on the fact that similar terms tend to be used in particular contexts. Consequently, they tend to coexist more. LSI is used to summarize text and retrieve and search for information based on the SVD technique. Another technique is LDA. LDA is a generative probabilistic model for discrete data collections such as text corpora. It is a three-level hierarchical Bayesian model in which each element of a collection is modeled as a finite mixture over an underlying set of topics. In turn, each theme is modeled as an infinite mixture over an underlying set of thematic probabilities. In the context of text modeling, topic probabilities provide an explicit representation of a document [31]. In the task of modeling a document, LDA does better than LSI and a mixture of unigram models. The LSI over-adjusts the probabilities of document modeling to determine the topics in a new document [31]. LDA generates less word ambiguity and a more accurate assignment of documents to topics. The LDA algorithm is easier to scale for large datasets using the MapReduce approach in a computational cluster. Therefore, to model a document, LDA is better than the LSI technique. The latter adjusts the probabilities of document modeling to determine the topics in a new document. The LDA model has proven to be a robust method that can be used to assign a class or category to an object to generate learning by creating a distinction between classes and cohesion within the same class. This study applies an LDA algorithm for topic modeling.

In the same methodological approach, Python libraries for clustering analysis were applied to the analysis of scientific production. The clustering technique is possibly the best-known unsupervised ML method [32], in which each element of a dataset is assigned to one or more automatically identified clusters. The clustering technique has been used, for example, to group the category results from web searches or for hierarchical clustering, where some clusters are contained within other clusters listed as higher-level concepts [33]. Some clustering algorithm approaches make it possible to characterize each group in relation to domains of interest. In this sense, groups can be named automatically, allowing people to understand similar elements within a given group. In fact, a text clustering resulting from an automated system analyzes the distribution of terms (words) in a body of text (e.g., titles and abstracts) and identifies groups of documents that use similar combinations of words; clustering "engines" often apply a descriptive term to each group to aid human interpretation [34,35]. The usefulness of each cluster label may vary depending on the algorithm's approach: data-focused algorithms focus on clustering text, and this is where K-means methods that vectorize the text contained in BOW are common.

BOW is an engineering model of typical features based on counting for textual data. This model may present accuracy problems when working with large corpora because the feature vectors are adjusted in absolute frequencies in absolute terms, which may lead to certain terms having frequencies—in all documents—causing terms without such high frequencies to disappear in the total feature set [36]. The term frequency inverse document frequency (TF-IDF) model lessens this problem. TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is a well-known algorithm that uses an input of a set of words that is commonly used to allocate a weight each word in the text document according to its uniqueness. In other words, the TF-IDF approach captures the relevance of particular words, text documents, and categories [37].

$$tfidf = tf * idf$$

where *tf* = term frequency; *idf* = inverse document frequency.

The term frequency *tf* in any document vector is denoted by the value of the raw frequency of that term in a particular document, represented as follows:

$$tf(w, D)\ fw_D$$

where fw_D = frequency of word w in document D, which becomes the term frequency (*tf*). Further,

$$idf(w, D) = 1 + log\frac{N}{1 + df(w)}$$

where *idf(w,D)* represents the *idf* of term/word w in document D, and N is the total number of documents in the corpus. *df (t)* is the number of documents in which the term w is present. Then, mathematically, the feature vector *tfidf* is represented as

$$tfidf = \frac{tfidf}{||tfidf||}$$

Consequently, this algorithm uses as input a set of words (representing a text). This is the output of the first text processing steps. The vectorization step transforms the words into a meaningful representation of numbers that is used to adjust the machine algorithm for prediction (the clustering algorithm in our case). We proceed to apply k-means, a technique for clustering data—an unsupervised ML technique [36]. This algorithm can classify unlabeled data into a predetermined number of groups based on similarities (k). The strengths of choosing this algorithm in this study are: (a) we had more than 1000 records, and k-means adapts well to large datasets, so it was considered a good choice from this perspective; (b) this type of algorithm guarantees convergence, so the result should be more robust clusters; and (c) it can group different shapes and sizes or data points or records. One of the main challenges of this algorithm is defining the best k. To handle this problem, we used the elbow method, which analyzes the percentage of variance explained as a function of the number of clusters. The first few groups will add much information, but at some point, the marginal gain will drop dramatically and result in an angle on the graph. The true "k," that is, the number of groups, is chosen at this point—hence the "elbow criterion" [38].

Considering the related works in this field of study, a thorough review of recent studies was carried out in which the methodological approach applied was analyzed. As can be seen in Table 1, similar methodologies were used that considered different time periods of scientific production in the area of competencies related to organizational leadership and management and BI&A. In all these studies, a metric analysis was applied, specifically bibliometric, although none of them incorporated scientometric or informetric analysis with advanced machine learning techniques to extract relevant data. However, the methodological approach designed for this study differs significantly from other studies in the following aspects:

Firstly, a database of indexed journals on scientific production in leadership competencies in BI&A over the last 20 years was compiled.

A two-stage methodological design is established: Stage 1 incorporates informetric analysis. This type of analysis is considered methodologically broader than bibliometric and scientometric studies. This strengthens the possibilities of analyzing phenomena and processes of different kinds, applying quantitative methods, and presenting descriptive characteristics, such as level of productivity in publications, scientific collaboration, and thematic structure, among many other characteristics [39].

Stage 2 focuses on deep learning techniques—types of artificial intelligence (AI)—methodologically in algorithms using a Python library for natural language processing, and therefore, applying deep learning techniques for the extraction of textual aspects of scientific development capable of recognizing and generating main topics with accurate prediction, in this case a technique called latent Dirichlet allocation (LDA). Analysis is incorporated with text clustering techniques, also belonging to the set of unsupervised algorithms to find groupings based on similarities that allow one to generate correlations to find the main clusters of scientific production in leadership competencies in BI&A of the last 20 years.

The results of these two methodological approaches to analysis (stage 1 and stage 2) were integrated into the analysis to validate and analyze consistency of results of the metric analysis of informetrics with the results of the associated IA methodology.

It is observed in Table 1 that the studies apply part of the metric analysis, mainly in bibliometrics. It is noted that the only study "Detection of emerging technologies and their evolution through deep learning and weak signal analysis" by Ebadi [40] methodologically used deep learning analysis techniques (not those of this study; there is no single type of algorithm that is best for solving a problem). However, none of the studies integrated the vision of informetrics with deep learning techniques, which are forms of artificial intelligence (AI).

Regarding the types of supervised machine learning and supervised non-machine learning algorithms and their evaluation metrics, a thorough classification study was performed for the most widespread techniques. For this purpose, we relied on studies with applications to different fields of recent scientific production.

Table 2 shows columns 1 (supervised algorithms) and 2 (literature sources), and a third column presenting the main evaluation metrics. The first column shows a list of important and generalized supervised machine learning algorithms—algorithms that need external assistance. For these, the input dataset is divided into train and test datasets. The train dataset has an output variable to be predicted or classified. The second column, upper part, indicates recent studies with applications of these algorithms. The lower part details the main metrics to evaluate these supervised algorithms—among these are sensitivity and specificity, number of leaves, number of decision variables, the confusion matrix, ROC curve, AUCPR, R-squared, root mean squared error (RMSE), and mean average precision (MAP), among others.

Column 3 (unsupervised algorithms) and column 4 (bibliographic sources) are now described. At the top of column 3, there is a list of important and generalized algorithms called unsupervised learning algorithms. Unlike supervised learning above, there are no correct answers, and there is no master. The algorithms are left to their own devices to discover and present interesting structures in the data. Unsupervised learning algorithms learn few features from the data. When new data are introduced, it uses previously learned features to recognize the class of the data. It is mainly used for clustering and feature reduction. The upper part of column 4 indicates sources of recent studies with applications of these algorithms. The lower part details the main metrics to evaluate these supervised algorithms; among these are scaling of variables; proportion of variance explained; optimal number of principal components; correlation of semantic terms; perplexity; coherence; elbow method among others.

This study's methodology uses unsupervised learning techniques and metrics of column 3 and 4—specifically a latent Dirichlet approach (LDA) and k-means clustering.

Table 1. Analysis of methodologies applied in the study on leadership competencies with BI&A, own elaboration.

		Data Base					**Informetric**		**IA**				**Validation**
		Years	**WoS**	**GS**	**Scopus**	**N° Articles**	**Bibliometric**	**Scientome**	**PLN**	**Lematizati**	**Clustering**	**AP**	**I +IA**
Data intelligence and analytics: A bibliometric analysis of human–Artificial intelligence in public sector decision-making effectiveness	Di Vaio, A. (2022) [25]	**2007–2021**	**x**		**x**	**161**	**x**						
Understanding the structure, characteristics, and future of collective intelligence using local and global bibliometric analyses	Calof, J. (2022) [41]	**1964–2004**	**x**		**x**	**3.138**	**x**						
Business Intelligence in Balanced Scorecard:Bibliometric analysis	Żółtowski, D. (2022) [42]	*	**x**	**x**		**>10.000**	**x**						
Detection of emerging technologies and their evolution through deep learning and weak-signal analysis	Ebadi, A. (2022) [40]	**1985–2020**	**x**			**590**			**x**			**x**	
Big data analytics and machine learning: A retrospective overview and bibliometric analysis	Zhang, JZ. (2021) [43]	**2006–2020**			**x**	**2.160**	**x**						
Influential and determinant models in big data analytics research: a bibliometric analysis	Aboelmaged, M. (2020) [44]	**2013–2019**	**x**		**x**	**229**	**x**						

* Various periods of years

	Data Base					**Informetric**		**IA**				**Validation**
	Years	**WoS**	**GS**	**Scopus**	**N° Articles**	**Bibliometric**	**Scientome**	**PLN**	**Lematizati**	**Clustering**	**AP**	**I +IA**
Data Analysis and Domain Knowledge for Strategic Compe-tencies Using Business Intelligence and Analytics	**1999–2021**	**x**		**x**		**x**	**x**	**x**	**x**	**x**	**x**	**x**

Table 2. Approaches supervised and unsupervised algorithms with main metrics; own elaboration.

Supervised Algorithms	Sources	NON-Supervised Algorithms	Sources
Decision Tree	Mahesh, B. (2020) [45]	Principal Component Analysis (PCA)	Mahesh, B. (2020) [45]
Navie Bayes	Ullah, I. (2022) [46]	Probabilistic latent semantic indexing (PLSI)	Suominen, A. (2016) [47]
Support Vector Machine	Chen, L. (2022) [48]	Latent Semantic Indexing (LSI)	Farkhod, A. (2021) [49]
Linear regression	Mayilvahanan, KS. (2022) [50]	Latent Dirichlet approach (LDA)	Tseng, SC. (2022) [51]
Logistic Regression	Tiwari, S. (2022) [52]	K-Means Clustering	Montavon, G. (2022) [53]
Main Evaluation Metrics (Supervised Learning)		**Main Evaluation Metrics (NON-Supervised Learning)**	
Decision Tree	Predictive accuracy rate; Accuracy rate: Sensitivity and specificity; Number of leaves; Number of decision variables; The confusion matrix	Principal Component Analysis (PCA)	Scaling of variables; Proportion of variance explained; Optimal number of principal components
Navie Bayes	Retention method	Probabilistic latent semantic indexing (PLSI)	Conditional probability distribution
Support Vector Machine	F1-Score;Precision;Recall Breakeven Point (PRBEP)	Latent Semantic Indexing (LSI)	Correlation of semantic terms
Linear regression	The confusion matrix; Recall, F1-Score;Area under the curve (AUC)	Latent Dirichlet approach (LDA)	Perplexity; Coherence
Logistic Regression	ROC curve; AUCPR;R-squared; root mean squared error (RMSE);Mean average precision (MAP)	K-Means Clustering	Elbow method

Comparison and Selection of Models

By comparing and selecting models for this study, we can see that the type of algorithm used depends on the type of problem to be solved, the number of variables, the type of model that best fits, etc. We delve into investigating algorithms commonly used in machine learning (ML) that are applicable to our study in Table 2.

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of probably correlated variable observations into a set of linearly uncorrelated variable values called principal components. It reduces the dimension of the data in order to proceed with faster and simpler calculations. It explains the variance-covariance structure of a set of variables through linear combinations. It is often used as a dimensionality reduction technique. [45]. This approach was not selected for our analysis objective, so it was not incorporated into the study methodologically.

In relation to latent patterns in semantic text, which, if it is the focus of analysis, has been studied in the literature as a likely technique to apply, probabilistic latent semantic indexing (PLSI) is able to draw latent patterns in semantic text. PLSI models each word in a document as a sample mixture model, where the mixture components are multinomial random variables that can be viewed as topic representations for clustering documents based on term frequencies. PLSI models a probability distribution in which the observed term and document frequency variables are mediated by a hidden topic variable. The number of topics is a hyperparameter delivered by the user. The output of the trained algorithm for a given document is a list of probabilities that each of the unlabeled topics discovered during training belongs to the document. PLSI uses internal parameters that can be seen as equivalent to the values derived by standard LSI. However, the two algorithms are different. The basic component of PLSI is classification, whereas LSI is a standard one used primarily for dimensionality reduction. LSI uses linear algebra, and PLSI is an iterative process based on expectation maximization. When researching literature to select the most accurate technique—studies comparing LSI with LDA—it is claimed that LDA has a better statistical basis for defining the θ topic-document distribution, by allowing inferences on new documents based on previously estimated models, and avoids the problem of overfitting. Latent Dirichlet allocation (LDA) is a subtype of PLSI with better statistical support based on two additional hyperparameters that make the algorithm prefer solutions where documents have relatively fewer topics (hyperparameter α) and where topics are characterized by relatively few words (hyperparameter β). The hyperparameters have values between 0 and 1; lower values impose a greater restriction on the number of topics/words. The literature review showed that LDA produces better results; hence, our study was oriented towards using LDA topic modelling as the methodological approach for stage two [49,51].

For ML stage-two clustering analysis, as one of the most important and fundamental techniques in ML, clustering has been extensively studied and applied to multiple fields of research. Clustering is a type of unsupervised learning which groups similar data points into the same group. In terms of similarity, the most commonly used criterion is distance, and k-means (KM) is a typical algorithm for this criterion. Classical k-means distributes data points to k different groups using norm distance 12.

3. Description of the Problem

Given the rapid development of BI&A, we were motivated to understand the development of scientific production in order to discover possible behavioral patterns in scientific development associated with the strategic competencies necessary for those who lead organizational management processes with BI&A. We wanted to be able to determine in which areas of competence domains the greatest emphasis is placed in the studies, in order to understand whether the focus of scientific progress has been toward mastery of professional skills for business managers through acquisition of a set of technical and theoretical skills and knowledge on business and analytical intelligence acquired in formal education processes; the development of a learning environment associated with the establishment of an adequate working environment that allows peer support, identification and dissemination of good practices and active learning of technologies associated with business intelligence and analytics (BI&A); the mastery of a strategic vision that refers to the ability to think creatively about the future, emerging contexts, trends, key aspects and imaginings of different future scenarios with the purpose of determining their implications and possible outcomes in a global and holistic perspective; and finally, whether the emphasis has been on the competencies that are essential for the development of a strategic vision—those that make it possible to integrate BI&A skills into models that work autonomously. The skills and knowledge acquired in business intelligence and analytics in situations that merit them are key in order to solve problems that arise in any organization and thus achieve the attainment of organizational objectives [22].

It is therefore very relevant to provide evidence on the level of development of the scientific production regarding strategic leadership competencies in BI&A and their relationships with the competence domains, to determine possible gaps and the different emphases, which will make it possible to perform more precise studies. These will allow a thorough understanding of the professional competencies needed by all those who must manage BI&A in a context of vertiginous development.

Consequently, this study analyzes the development of the scientific production of the last 20 years in strategic leadership competencies for organizational management in the field of BI&A due to the demanding demand for the development of new competencies in BI&A techniques, for models and software for decision making by those who lead organizations.

This study sheds light on the state of the art of BI&A in the field of strategic competencies for organizational leadership, helping to inform readers that the dimensions of strategic leadership in BI&A form a fertile and largely unexplored field. At the same time, this study has contributed through the application of metrics and AI techniques to determining gaps in the field of BI&A and organizational strategic-leadership-competency research.

4. Methodology

4.1. Hypotheses

Building and applying a new vision for future strategy, based on the opportunity perceived with BI&A, is essential to developing a degree of self-knowledge to anticipate the future and make accurate strategic assessments of the organization's resources and capabilities. Intellectual agility is the ability to adapt, innovate, and transform ideas into new and improved products, processes and internal and external services, and it appears to be a critical factor related to BI&A [54]. CEOs, who lead organizations, require both an appreciation of and familiarity with BI&A, depending on the level of organizational development. It is not necessary to have a background in statistics, but such leaders must understand the theory behind various quantitative methods and the demand for extensive training for employees [8]. Skills are needed to assess technology and data infrastructures thoroughly and thus understand the technology gaps. This means having an analytics-driven business culture, i.e., the ability to identify, design and implement business use cases, through having data and technology capabilities appropriate for the data infrastructure, for the development and implementation of complex analytics; and by having capable individuals in the organization—quantitative professionals and organizational management specialists.

4.2. Methodological Steps

To determine the level of leadership development and BI&A, we designed a methodology that integrates analytical techniques of informetrics, complemented by natural language processing and machine learning techniques.

Indeed, the informetric technique makes it possible to show complex relationships between its various categories, such as documents, authors, research groups, journals, institutions, countries, and regions. This is achieved thanks to data mining, information processing, statistics, mathematics, visualization programs and bibliographic databases [55]. As a result of the combination of the above, informetrics allows for the construction of a more organized and understandable knowledge base [56].

Machine learning, on the other hand, is concerned with the formal study of learning systems. It is a highly interdisciplinary area of knowledge in which various theories, models and statistical tools from computer science, engineering, cognitive science, optimization theory and other disciplines of science and mathematics coexist. Among the different types of auto-mathematical learning, one can distinguish unsupervised machine learning. In this type of algorithm, the machine can learn, since it does not receive any feedback from its environment. It is possible to develop a formal framework for unsupervised learning based on the notion that the goal of the machine is to build representations of the input that can be used to make decisions, predict future inputs, efficiently communicate the inputs to another machine, etc. In a sense, unsupervised learning can be thought of as finding patterns in data beyond what would be considered pure unstructured noise [57].

In that direction, we developed two models from the AI domain whose approaches are based on unsupervised machine learning. The first one uses topic-modelling algorithms, which are widely used and have proven to be successful in the area of opinion mining to extract "latent" topics that relate to aspects of interest. The second uses unsupervised machine learning algorithms for text clustering, which partition a set of clustered text documents based on distance or similarity measures [58]. Effectively, text clustering algorithms attempt to develop an appropriate clustering of large text documents based on well-formulated criteria represented by an objective function. The objective function is an equation for evaluating given constraints—with minimized or maximized objectives—using a non-linear programming system. It is used in the text-clustering domain to partition decisions to distribute a set of documents over a subset of cluster centers [59].

The two models—with an unsupervised machine learning approach in this research—work on their own to discover the inherent structure of the unlabeled data (which is the situation in unsupervised learning algorithms in general). Thus, we followed the evaluation methods recommended for this type of problem: (1) we used consistent scores to check if the discovered knowledge is consistent; (2) an expert validated that the results make sense; see Figure 1.

1. Topic-modelling algorithms have proven to be successful in the area of aspect-based opinion mining to extract "latent" topics, which are aspects of interest. A technique called latent Dirichlet allocation (LDA) is used, which is based on a generative probabilistic model in which each document consists of a combination of several topics, where terms or words can be assigned to a specific topic. Latent LDA is a good-topic-modeling algorithm compared to latent semantic analysis and the hierarchical Dirichlet process for the aspect extraction process in aspect-based opinion mining [60]. The results of this technique were used to analyze the most relevant topics of the scientific production regarding strategic leadership competencies in BI&A and their relationships with the competence domains.
2. The second unsupervised machine learning algorithm was applied to a grouping of texts in order to analyze the main clusters resulting from the 1231 articles considered. The aim was to analyze the results of the k-means model trained to predict the type of cluster belonging to each article related to each of these and to analyze the resulting pattern of the most relevant scientific production with respect to the emphasis on the dimension of strategic leadership competencies in BI&A and its relationship with the competence domains.

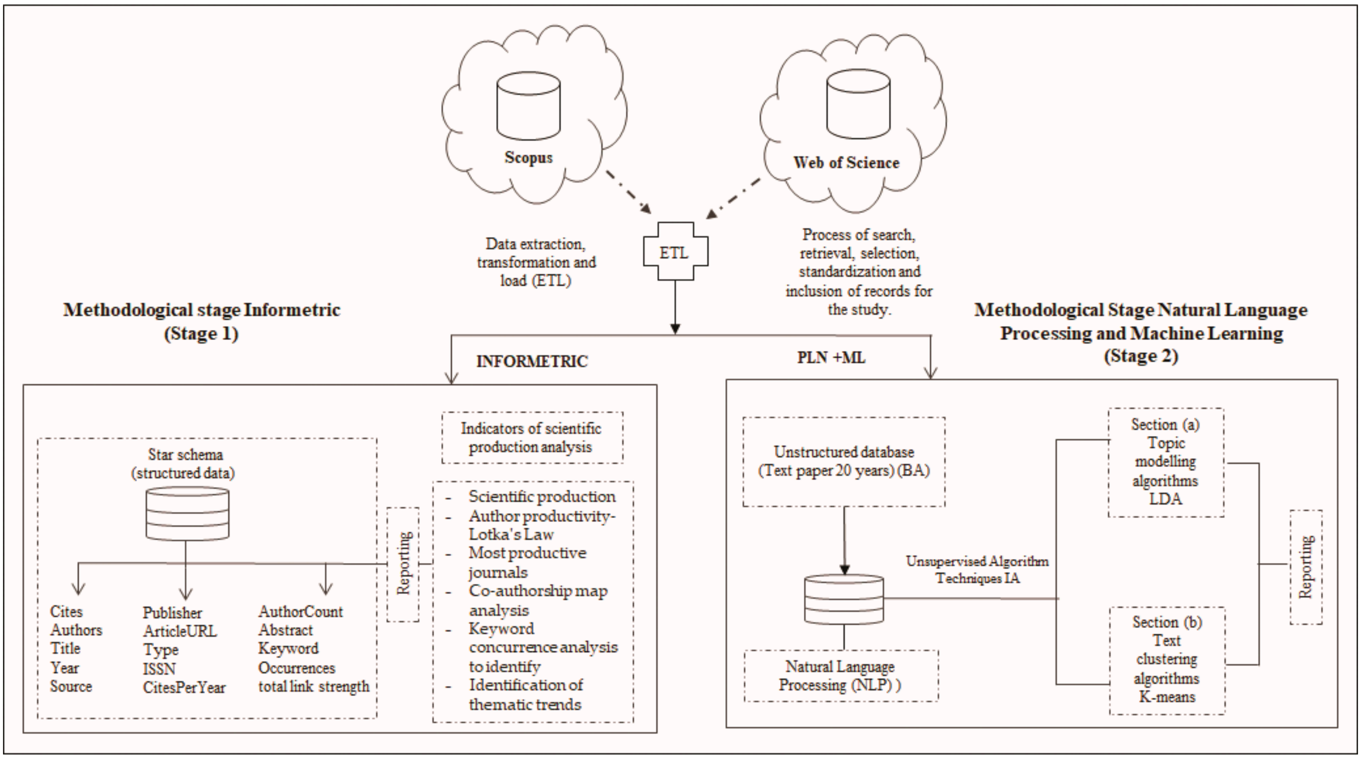

Scopus
Web of Science
Data extraction, transformation and load (ETL)
ETL
Process of search, retrieval, selection, standardization and inclusion of records for the study.
Methodological stage Informetric (Stage 1)
Methodological Stage Natural Language Processing and Machine Learning (Stage 2)
INFORMETRIC
PLN +ML
Star schema (structured data)
Cites
Authors
Title
Year
Source
Publisher
ArticleURL
Type
ISSN
CitesPerYear
AuthorCount
Abstract
Keyword
Occurrences
total link strength
Reporting
Indicators of scientific production analysis
- Scientific production
- Author productivity-Lotka's Law
- Most productive journals
- Co-authorship map analysis
- Keyword concurrence analysis to identify
- Identification of thematic trends
Unstructured database (Text paper 20 years) (BA)
Unsupervised Algorithm Techniques IA
Natural Language Processing (NLP))
Section (a) Topic modelling algorithms LDA
Section (b) Text clustering algorithms K-means
Reporting

Figure 1. Methodological design for information analysis prepared by the authors.

Research Data Sample:

For the informetric methodological stage (stage 1) and the natural language processing and machine learning methodological stage (stage 2), the data sample was obtained by means of a search protocol (Table 3) from the most recognized multidisciplinary database platforms worldwide, Web of Science (WoS) Core Collection and Scopus, without a time restriction. The document typologies were "article", "review", "conference paper" and proceedings papers. The 350 and 1058 records retrieved from Web of Science and Scopus (Figure 2) were exported to EndNote, which enabled the elimination of duplicates and the subsequent creation of a single database with 1231 documents. Quality control of the documents was then carried out in accordance with the objectives of the study. The date of extraction of records: 16-08-21.

Table 3. Advanced search equation for each bibliographic database.

Web of Science	Scopus
*TS = ("business intelligence" AND (analytical OR strategic OR analysis OR descriptive OR predictive OR prescriptive OR competitive OR "Analytics 1.0" OR "Analytics 2.0" OR "Analytics 3.0" OR "Analytics 4.0") AND (Leadership OR Models OR Competenc * s OR "competency center" OR leadership OR capability * OR skill * OR ability *) AND ("big data" OR "data warehouse" OR "machine learning" OR "predictive modeling" OR mobile OR dashboard OR cloud OR "data mining" OR "Artificial Intelligence" OR OLAP) AND (exploratory OR benefits OR implementation OR solutions OR success OR satisfaction OR decision OR continuum OR management OR adoption OR benefits OR implementation))* *Refined by: DOCUMENT TYPES: (Article OR review OR proceedings papers)* Indexes: SCI-EXPANDED, SSCI, A&HCI, ESCI.	*TITLE-ABS-KEY ("business intelligence" AND (analytical OR strategic OR analysis OR descriptive OR predictive OR prescriptive OR competitive OR "Analytics 1.0" OR "Analytics 2.0" OR "Analytics 3.0" OR "Analytics 4.0") AND (leadership OR models OR competenc * s OR "competency center" OR leadership OR capability * OR skill * OR ability *) AND ("big data" OR "data warehouse" OR "machine learning" OR "predictive modeling" OR mobile OR dashboard OR cloud OR "data mining" OR "Artificial Intelligence" OR olap) AND (exploratory OR benefits OR implementation OR solutions OR success OR satisfaction OR decision OR continuum OR management OR adoption OR benefits OR implementation)) AND (LIMIT-TO (DOCTYPE, "cp") OR LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "re"))*

*(Metacharacter *: Represents characters or ranges of previous values that can be matched zero or more times).*

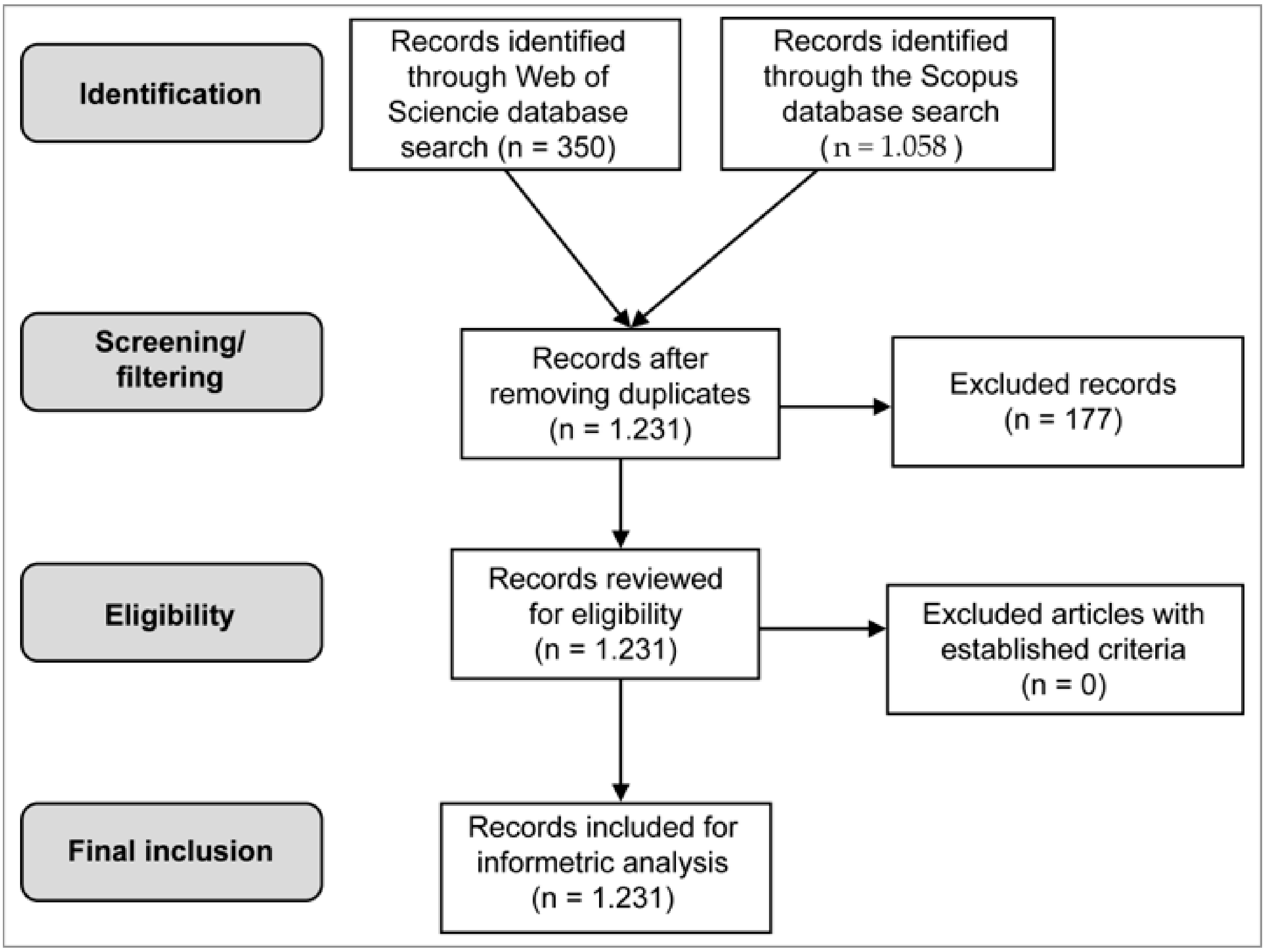

Identification
Records identified through Web of Sciencie database search (n = 350)
Records identified through the Scopus database search (n = 1.058)
Screening/ filtering
Records after removing duplicates (n = 1.231)
Excluded records (n = 177)
Eligibility
Records reviewed for eligibility (n = 1.231)
Excluded articles with established criteria (n = 0)
Final inclusion
Records included for informetric analysis (n = 1.231)

Figure 2. Methodological process for search, collection, selection, normalization and inclusion of records for the study.

Accordingly, the methodology of this research consists of two stages, as described below:

The informetric methodological stage (stage 1) addressed the information of the descriptive study and examined the scientific production on competencies, professional skills in the field of knowledge in business intelligence (BI) and analytical intelligence (BA) in organizations. To achieve this purpose, the evolution of scientific production, the productivity of authors according to Lotka's law, the most productive journals, the co-authorship map and the co-occurrence of keywords were analyzed to identify thematic trends (Figure 1).

The natural language processing and machine learning stage (stage 2) considered the analysis from the point of view of content, specifically the titles and abstracts of scientific production, using natural language processing and machine learning techniques. Indeed, the methodological idea was to extract key themes or concepts from a corpus of documents generated by NPL from the last 20 years of publications in the area of study of this article, and to represent them as topics. Each theme is represented as a bag or collection of words (or terms) from the document corpus. The latent Dirichlet allocation (LDA) technique was used, which is based on a generative probabilistic model in which each document consists of a combination of several topics; terms or words can be assigned to a specific topic. As a second additional analysis, the text clustering technique was applied to find the main clusters in the 1231 articles considered. The methodological scheme is in Figure 1.

Application of the Informetrics Methodology (Stage 1)

By means of a search protocol, information was retrieved from the multidisciplinary database platforms with the greatest worldwide recognition, Web of Science (WoS) Core Collection and Scopus, without a time restriction, as the behavior of the subject from its beginnings in the aforementioned databases was analyzed, as can be seen in Figure 2. Prior to an exhaustive review of the literature, the search equations described for each bibliographic database (Table 3) were used to retrieve records and create the dataset for the study, during the record extraction: 16-08-21.

The TS and TITLE-ABS-KEY field codes retrieved the records used, including the titles, abstracts and keywords of the documents. The document typologies were "article", "review", "conference paper" or proceedings papers. The 350 and 1058 records retrieved from Web of Science and Scopus, respectively (see Figure 2), were exported to EndNote, which enabled duplicates to be removed and the subsequent single database with 1231 documents to be created. Quality control of the documents was then carried out in accordance with the objectives of the study.

From the records retrieved, the units of analysis considered were articles, authors, documents and keywords. The units of measurement were indicators of productivity, collaboration and keyword co-occurrence using bibliometric network maps. To analyze the information, Excel 2019, Publish or Perish 8, EndNote X9, Tableau 2022 and VOSviewer v1.6.18 were used. VOSviewer was used to map distance-based bibliometric networks [61]. Keyword co-occurrence is the measure of the frequency of keyword pairs in a set of documents that serve to detect sources, trends or the scientific structure of research fields [62,63]. Clusters in the network are represented by specific colors and are made up of a set of nodes or items closely related to each other, according to the co-occurrence of keywords, where each node is assigned to only one cluster (Table 4).

Table 4. Description of bibliometric indicators and analysis tools.

Analyzed Dimensions	Indicators/Variables: Description
	Scientific Production
Scientific activity	• Number of texts: Calculated by year and by type • Lotka's law of productivity of authors • Productivity by scientific journal
Scientific collaboration	• Collaboration index • Degree of collaboration • Collaboration coefficient
Structural analysis	Statistical technique and variables
Thematic structure	Keyword co-occurrence network

4.3. Natural Language Processing and Machine Learning—Stage 2

Machine learning (ML) can be understood as a sub-area of AI. ML allows machines to learn automatically, gaining patterns and insights from data. It is common to use a combination of ML and NLP to solve problems such as text categorization, parsing unstructured text data into more structured forms and clustering. [36]. The stage-2 analysis comprised a collection of ML, linguistic and statistical techniques that were used to model and extract information from text primarily for analysis purposes. This analysis was based on unstructured data sources of 1231 articles, in PDF format, corresponding to all the abstracts and titles of articles on which NLP techniques were applied. The most technical tools and algorithms will be used efficiently to understand the unstructured text data of scientific production. The methodological idea is to extract key themes or concepts from a corpus of documents generated by NLP between 1999 and 2021 among scientific publications in the area of study. Each topic can be represented as a bag or collection of words (or terms) from the document corpus. We used the LDA technique, based on a generative probabilistic model, where documents constitute a combination of several topics, and where terms or words can be assigned to a specific topic [31].

4.3.1. Natural Language Processing (NLP) Techniques

Text processing techniques

(1) **Analysis**

To represent the main themes, during the first data analysis, textual columns were extracted: abstract and title for each article. Pre-processing (so-called text pre-processing, i.e., removing noisy terms and data) of unstructured texts was carried out.

(2) **Tokenization**

Therefore, a wide variety of techniques are applied that convert plain text into well-defined sequences of linguistic components that have a standard structure and notation. Therefore, each row of data was converted into a list of lowercase tokens (the NLTK tool was used, which is a platform for Python programming, providing several interfaces to perform sentence tokenization). This transformation was performed using the gensim tool (simple_preprocess). The result of this step is represented in the following example:

['apply', 'big', 'data', 'analysis', 'in', 'higher', 'education'].

(3) **Construction of bigram and trigram models**

Then, the bigram and trigram models were constructed. With gensim, common phrases, i.e., multi-word expressions, n-gram collocations of words from text sentences, were detected. All words and bigrams with a total collected count of less than 5 (called min_count) were ignored. This tool generates texts (n-grams) that are highly connected to each other; i.e., it accepts n-grams that meet this condition: (count(a, b)-min_count) * N/(count(a) * count(b)) > threshold, where N is the total vocabulary size in all data.

Experimentally, it was found that threshold = 100 is good enough to extract adequate and consistent n-grams. An example of word outputs is shown below:

['applying', 'big', 'data', 'analytics', 'in', 'higher_education', 'systematic',
'mapping', 'study', 'higher_education', 'systems', 'hes', 'have', 'become',
'increasingly', 'absorbed', 'in', 'applying', 'data', 'analytics' . . .]

(4) **Applying natural language processing**

(a) **Elimination of empty words:**

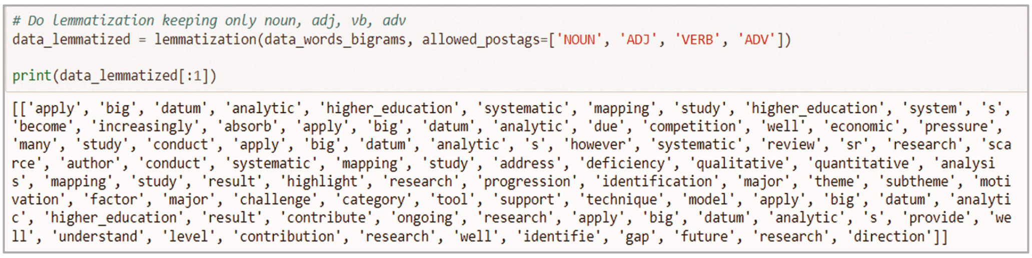
The elimination of empty words in English that do not add much meaning to a sentence was continued.

(b) **Lemmatization:**

They can be safely ignored without sacrificing the meaning of the sentence. We proceeded to the lemmatization of the generated n-grams. In this step, the inflected forms of a word were grouped together so that they could be analyzed as a single element, identified by the word's lemma or dictionary form.

(c) **POS tagging:**

POS tagging was established. The syntax of the n-grams generated by categorizing words in a text (corpus) in correspondence with a particular part of speech was analyzed, depending on the definition of the word and its context, and then they were filtered by tags, while keeping the works with the tag ['NOUN', 'ADJ', 'VERB', 'ADV'], as they describe the main concepts and actions in articles, see code 1 in Figure 3.



```
# Do lemmatization keeping only noun, adj, vb, adv
data_lemmatized = lemmatization(data_words_bigrams, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV'])

print(data_lemmatized[:1])
```

```
[['apply', 'big', 'datum', 'analytic', 'higher_education', 'systematic', 'mapping', 'study', 'higher_education', 'system', 's',
'become', 'increasingly', 'absorb', 'apply', 'big', 'datum', 'analytic', 'due', 'competition', 'well', 'economic', 'pressure',
'many', 'study', 'conduct', 'apply', 'big', 'datum', 'analytic', 's', 'however', 'systematic', 'review', 'sr', 'research', 'sca
rce', 'author', 'conduct', 'systematic', 'mapping', 'study', 'address', 'deficiency', 'qualitative', 'quantitative', 'analysi
s', 'mapping', 'study', 'result', 'highlight', 'research', 'progression', 'identification', 'major', 'theme', 'subtheme', 'moti
vation', 'factor', 'major', 'challenge', 'category', 'tool', 'support', 'technique', 'model', 'apply', 'big', 'datum', 'analyti
c', 'higher_education', 'result', 'contribute', 'ongoing', 'research', 'apply', 'big', 'datum', 'analytic', 's', 'provide', 'we
ll', 'understand', 'level', 'contribution', 'research', 'well', 'identifie', 'gap', 'future', 'research', 'direction']]
```

Figure 3. Code 1: POS tagging, own elaboration.

The final processed text resulting from the last step's lemmatized words, all of them related to the defined set of accepted tags, was used as input for the following analyses.

4.3.2. Application of the Natural Language Processing and Machine Learning (Stage 2)

4.3.3. Section (a) Topic Modelling Algorithms: LDA Model

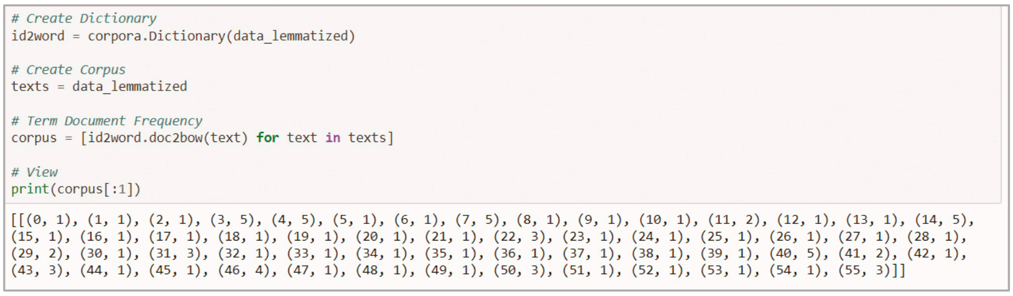
Recognizing main topics

(1) **Dictionary word assignment**

We proceeded to map words with dictionary IDs which encapsulate the mapping between normalized words (the output of the last stage) and their integer ids.

(2) **Construction of BOW representations**

Next, BOW (or bag of words) representations were constructed. In this process, each item in the dataset was converted (changed to a list of words and n-grams) to bag-of-words format, i.e., a list of pairs in the format (token_id, token_count). This representation describes the occurrence of words within a document. There are two parts in it: token_id, which refers to a vocabulary of known words, and token_count, which reflects a measure of the presence of known words. An example output of this form of representation is shown as follows: [(0, 1), (1, 1), (2, 1), (3, 5), (4, 5), (5, 1), (6, 1), (7, 5) . . .]. This representation shows that the word with id = 0 appears once in the first document (because we have (0,1) in the result) and so on, see Figure 4 below.



```
# Create Dictionary
id2word = corpora.Dictionary(data_lemmatized)

# Create Corpus
texts = data_lemmatized

# Term Document Frequency
corpus = [id2word.doc2bow(text) for text in texts]

# View
print(corpus[:1])
```

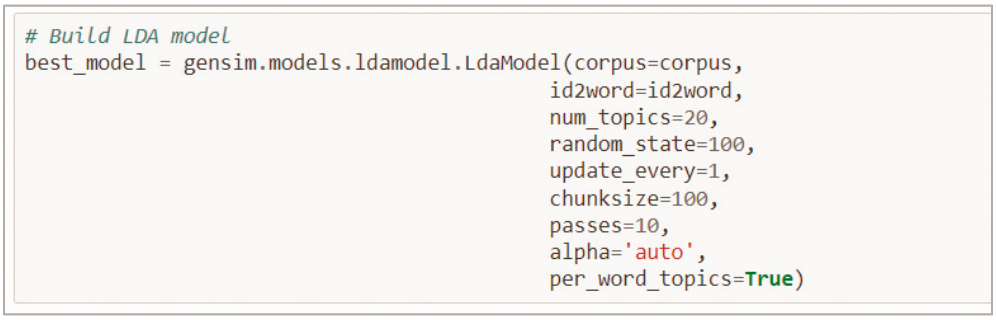
```
[[(0, 1), (1, 1), (2, 1), (3, 5), (4, 5), (5, 1), (6, 1), (7, 5), (8, 1), (9, 1), (10, 1), (11, 2), (12, 1), (13, 1), (14, 5),
(15, 1), (16, 1), (17, 1), (18, 1), (19, 1), (20, 1), (21, 1), (22, 3), (23, 1), (24, 1), (25, 1), (26, 1), (27, 1), (28, 1),
(29, 2), (30, 1), (31, 3), (32, 1), (33, 1), (34, 1), (35, 1), (36, 1), (37, 1), (38, 1), (39, 1), (40, 5), (41, 2), (42, 1),
(43, 3), (44, 1), (45, 1), (46, 4), (47, 1), (48, 1), (49, 1), (50, 3), (51, 1), (52, 1), (53, 1), (54, 1), (55, 3)]]
```

Figure 4. Code 2: Recognizing main topics, own elaboration.

4.3.4. Development of a Research Model, see code 3 in Figure 5

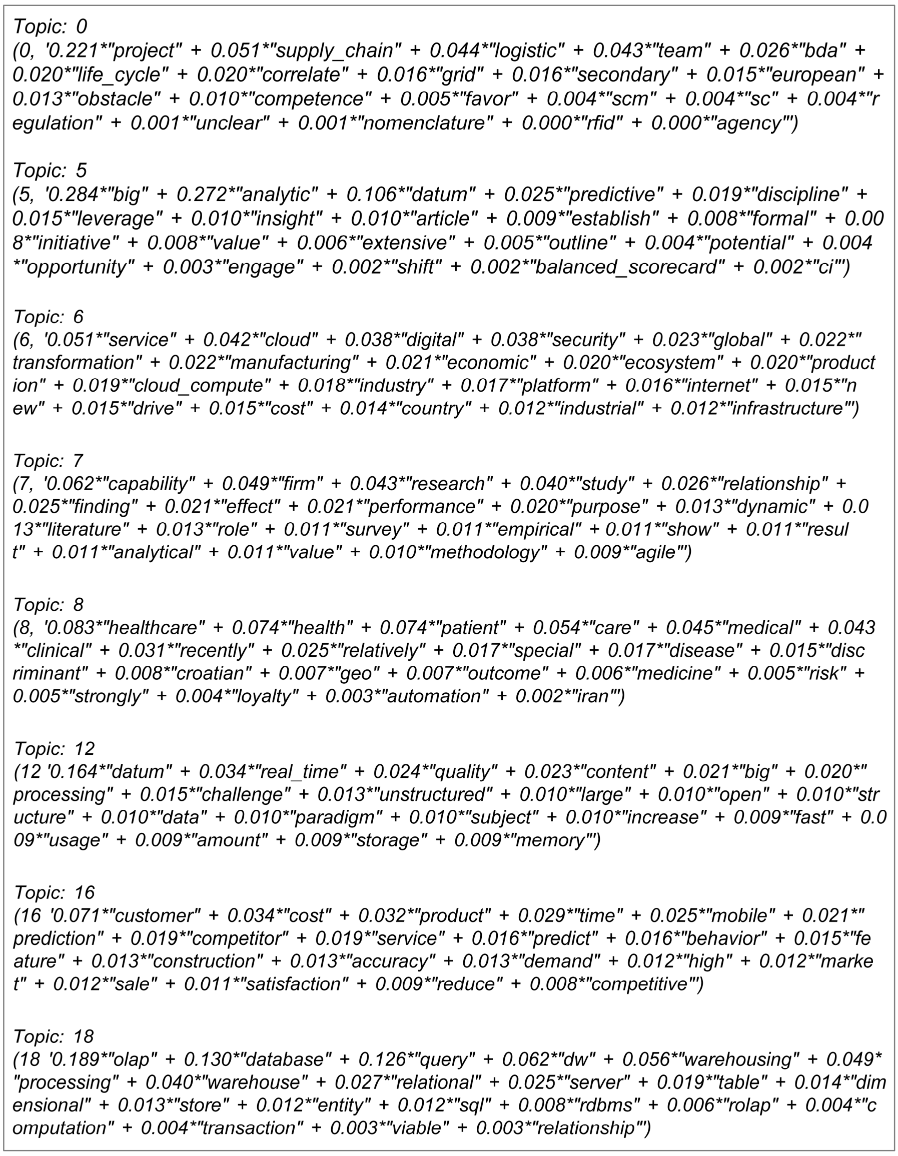
Therefore, on the basis of the last steps, we can recognize the resulting main themes (Figure 6).

To assess the quality of the generated topic models, perplexity and coherence scores were used as measures. From the training corpus, the quality was evaluated by calculating the probability assigned to the text strings of the test corpus. To avoid probabilities that would be too small, perplexity was used. In general, the lower the perplexity, the better the model [31,64]. To address possible limitations of the perplexity measure, human judgment was used to determine when the generated model produced topics with greater coherence in relation to semantically related terms, thereby identifying a global idea of the same topic. Thus, a clear level of identification was achieved considering the mastery of the concepts of the general context [64]. To define the ideal number of topics, we used topic coherence to compare the different possible k values (number of models) with each other, see code 4 in Figure 7.



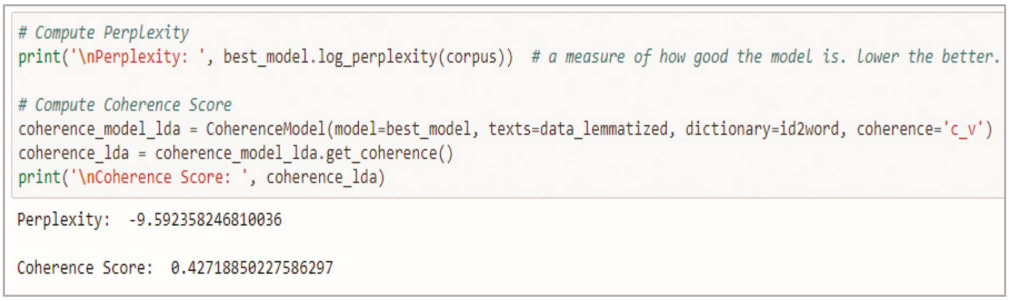
```
# Build LDA model
best_model = gensim.models.ldamodel.LdaModel(corpus=corpus,
                                             id2word=id2word,
                                             num_topics=20,
                                             random_state=100,
                                             update_every=1,
                                             chunksize=100,
                                             passes=10,
                                             alpha='auto',
                                             per_word_topics=True)
```

Figure 5. Code 3: Development of a Research Model, own elaboration.



```
Topic:  0
(0, '0.221*"project" + 0.051*"supply_chain" + 0.044*"logistic" + 0.043*"team" + 0.026*"bda" +
0.020*"life_cycle" + 0.020*"correlate" + 0.016*"grid" + 0.016*"secondary" + 0.015*"european" +
0.013*"obstacle" + 0.010*"competence" + 0.005*"favor" + 0.004*"scm" + 0.004*"sc" + 0.004*"r
egulation" + 0.001*"unclear" + 0.001*"nomenclature" + 0.000*"rfid" + 0.000*"agency"')

Topic:  5
(5, '0.284*"big" + 0.272*"analytic" + 0.106*"datum" + 0.025*"predictive" + 0.019*"discipline" +
0.015*"leverage" + 0.010*"insight" + 0.010*"article" + 0.009*"establish" + 0.008*"formal" + 0.00
8*"initiative" + 0.008*"value" + 0.006*"extensive" + 0.005*"outline" + 0.004*"potential" + 0.004
*"opportunity" + 0.003*"engage" + 0.002*"shift" + 0.002*"balanced_scorecard" + 0.002*"ci"')

Topic:  6
(6, '0.051*"service" + 0.042*"cloud" + 0.038*"digital" + 0.038*"security" + 0.023*"global" + 0.022*"
transformation" + 0.022*"manufacturing" + 0.021*"economic" + 0.020*"ecosystem" + 0.020*"product
ion" + 0.019*"cloud_compute" + 0.018*"industry" + 0.017*"platform" + 0.016*"internet" + 0.015*"n
ew" + 0.015*"drive" + 0.015*"cost" + 0.014*"country" + 0.012*"industrial" + 0.012*"infrastructure"')

Topic:  7
(7, '0.062*"capability" + 0.049*"firm" + 0.043*"research" + 0.040*"study" + 0.026*"relationship" +
0.025*"finding" + 0.021*"effect" + 0.021*"performance" + 0.020*"purpose" + 0.013*"dynamic" + 0.0
13*"literature" + 0.013*"role" + 0.011*"survey" + 0.011*"empirical" + 0.011*"show" + 0.011*"resul
t" + 0.011*"analytical" + 0.011*"value" + 0.010*"methodology" + 0.009*"agile"')

Topic:  8
(8, '0.083*"healthcare" + 0.074*"health" + 0.074*"patient" + 0.054*"care" + 0.045*"medical" + 0.043
*"clinical" + 0.031*"recently" + 0.025*"relatively" + 0.017*"special" + 0.017*"disease" + 0.015*"disc
riminant" + 0.008*"croatian" + 0.007*"geo" + 0.007*"outcome" + 0.006*"medicine" + 0.005*"risk" +
0.005*"strongly" + 0.004*"loyalty" + 0.003*"automation" + 0.002*"iran"')

Topic:  12
(12 '0.164*"datum" + 0.034*"real_time" + 0.024*"quality" + 0.023*"content" + 0.021*"big" + 0.020*"
processing" + 0.015*"challenge" + 0.013*"unstructured" + 0.010*"large" + 0.010*"open" + 0.010*"str
ucture" + 0.010*"data" + 0.010*"paradigm" + 0.010*"subject" + 0.010*"increase" + 0.009*"fast" + 0.0
09*"usage" + 0.009*"amount" + 0.009*"storage" + 0.009*"memory"')

Topic:  16
(16 '0.071*"customer" + 0.034*"cost" + 0.032*"product" + 0.029*"time" + 0.025*"mobile" + 0.021*"
prediction" + 0.019*"competitor" + 0.019*"service" + 0.016*"predict" + 0.016*"behavior" + 0.015*"fe
ature" + 0.013*"construction" + 0.013*"accuracy" + 0.013*"demand" + 0.012*"high" + 0.012*"marke
t" + 0.012*"sale" + 0.011*"satisfaction" + 0.009*"reduce" + 0.008*"competitive"')

Topic:  18
(18 '0.189*"olap" + 0.130*"database" + 0.126*"query" + 0.062*"dw" + 0.056*"warehousing" + 0.049*
"processing" + 0.040*"warehouse" + 0.027*"relational" + 0.025*"server" + 0.019*"table" + 0.014*"dim
ensional" + 0.013*"store" + 0.012*"entity" + 0.012*"sql" + 0.008*"rdbms" + 0.006*"rolap" + 0.004*"c
omputation" + 0.004*"transaction" + 0.003*"viable" + 0.003*"relationship"')
```

Figure 6. Selected major topics discovered.



```
# Compute Perplexity
print('\nPerplexity: ', best_model.log_perplexity(corpus))  # a measure of how good the model is. lower the better.

# Compute Coherence Score
coherence_model_lda = CoherenceModel(model=best_model, texts=data_lemmatized, dictionary=id2word, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score: ', coherence_lda)
```

```
Perplexity:  -9.592358246810036

Coherence Score:  0.42718850227586297
```

Figure 7. Code 4: metrics perplexity and coherence, own elaboration.

Looking at the perplexity calculation, a good UMass score value of −9.592358246810036 was obtained. Similarly, the higher the Cv score in coherence, the better the model [36,65]. From the iteration process, the number of topics with the highest coherence score (0.42718850227586297, equivalent to 20) was selected as the optimal value. Similarly, contribution (proportion) tests were performed for the most significant topics. We selected 8 of the 20 topics with the highest contribution and excluded those that did not represent a trend in topics that could be related to a generic domain, such as “software” or “computers”. Both were considered outliers and were excluded based on these two heuristics.

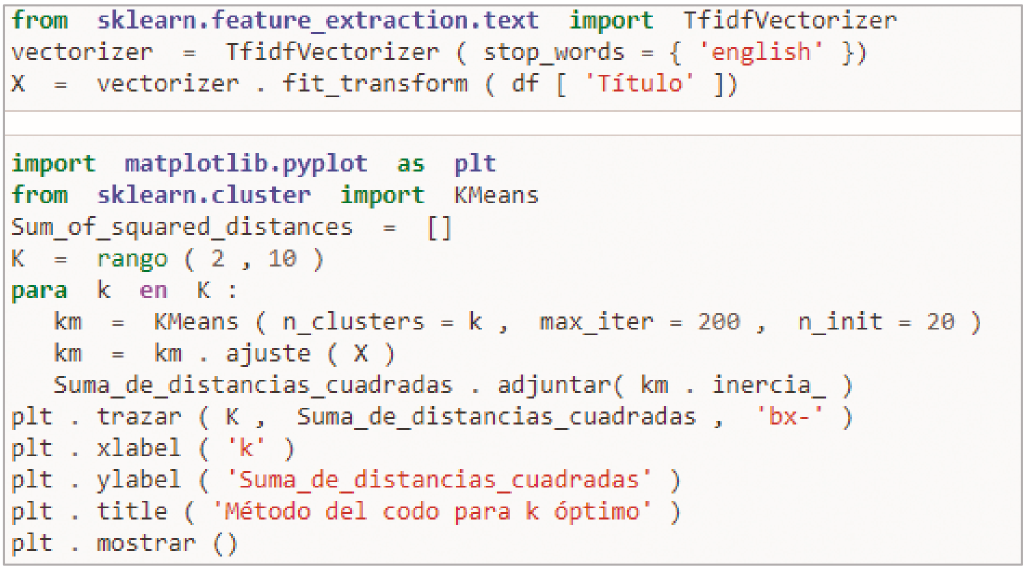
4.3.5. Section (b) Text Clustering Algorithms

Research model development

1. **TF-IDF Vectorization:**

TF-IDF stands for term frequency inverse document frequency. It takes as input a set of words (representing a text); in our case, it is the output of the first steps of text processing.

The vectorization step transforms the words into a meaningful representation of numbers that is used to adjust the algorithm of the prediction machine (the clustering algorithm in our case). The final result of this step is the TF-IDF of each word that appears in the whole text. TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates the relevance of a word in a document in a document collection, see code 5 in Figure 8.



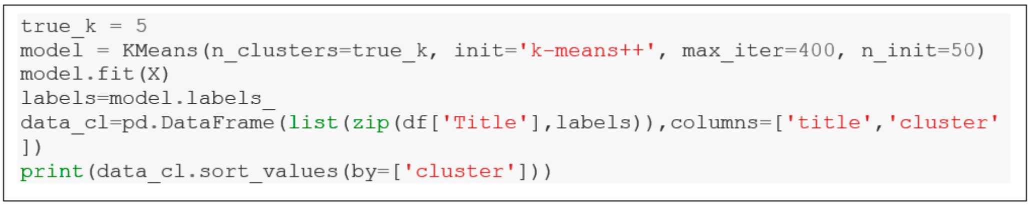
```
from  sklearn.feature_extraction.text  import  TfidfVectorizer
vectorizer  =  TfidfVectorizer ( stop_words = { 'english' })
X  =  vectorizer . fit_transform ( df [ 'Título' ])
```

```
import  matplotlib.pyplot  as  plt
from  sklearn.cluster  import  KMeans
Sum_of_squared_distances  =  []
K  =  rango ( 2 , 10 )
para  k  en  K :
   km  =  KMeans ( n_clusters = k ,  max_iter = 200 ,  n_init = 20 )
   km  =  km . ajuste ( X )
   Suma_de_distancias_cuadradas . adjuntar( km . inercia_ )
plt . trazar ( K ,  Suma_de_distancias_cuadradas ,  'bx-' )
plt . xlabel ( 'k' )
plt . ylabel ( 'Suma_de_distancias_cuadradas' )
plt . title ( 'Método del codo para k óptimo' )
plt . mostrar ()
```

Figure 8. Code 5: TF-IDF Vectorization, own elaboration.

2. **Applying K-Means:**

K-means is a data clustering technique that can be used for unsupervised machine learning. It is able to classify unlabeled data into a predetermined number of clusters based on similarities (k). One of the main challenges of this algorithm is to define the best k. To solve this problem, we used the elbow method (next step), It is self-elaborated, not from another author, so it is not cited, the phrase, see code 6 in Figure 9.



```
true_k = 5
model = KMeans(n_clusters=true_k, init='k-means++', max_iter=400, n_init=50)
model.fit(X)
labels=model.labels_
data_cl=pd.DataFrame(list(zip(df['Title'],labels)),columns=['title','cluster'
])
print(data_cl.sort_values(by=['cluster']))
```

Figure 9. Code 6: Applying K-Means, own elaboration.

3. **Elbow method for optimal k**

The elbow method performs k-means clustering on a dataset for a range of k values (say 1 to 10). It consists of two main steps:

a. Performing k-means clustering with all these different values of k.

b. Plotting these points and finding the point where the mean distance to the centroid drops sharply ("elbow"), see Figure 10.

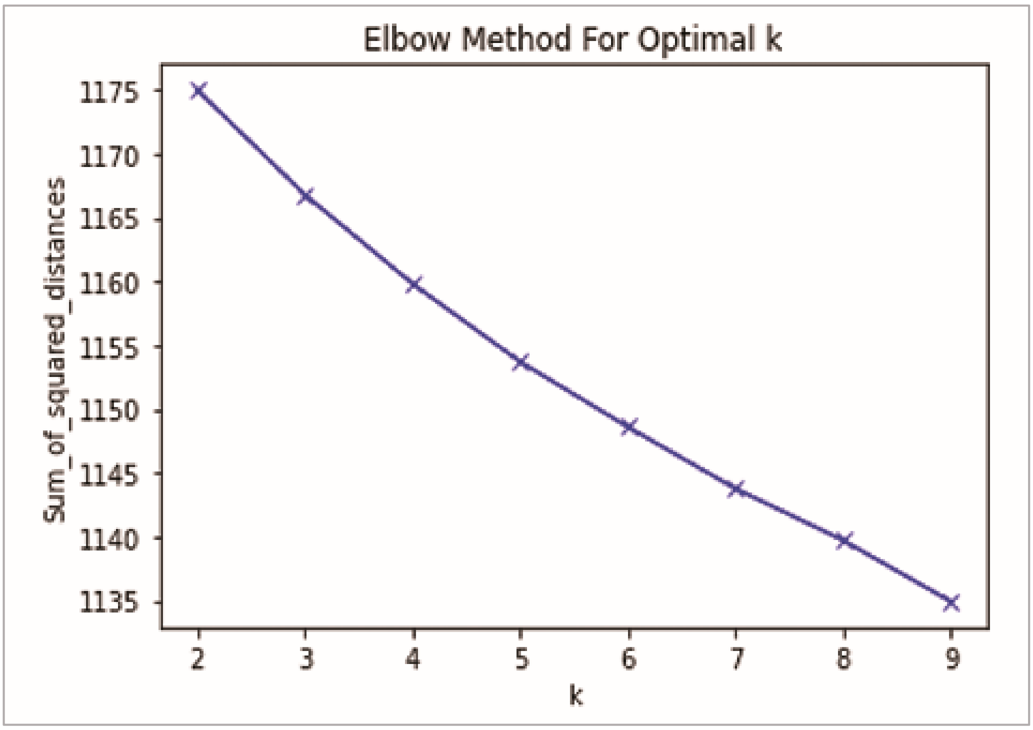


Figure 10. Elbow method for optimal k own elaboration.

5. Results

The results of this research correspond to the analysis of the informetric stage (stage 1) and the natural language processing (NLP) and machine learning (ML) stage (stage 2) and are presented below. No studies were found that integrated informetric analysis with other ML techniques to study scientific production in competencies in the field of BI&A and organizational leadership.

5.1. Results of Stage 1

5.1.1. Scientific Production According to Documentary Typology

The three types of documents analyzed (i.e., articles, reviews, and conference paper) had different behaviors in terms of the number of publications throughout the study period. We found 561 articles, 658 conference papers and 12 reviews. As for the articles, the greatest number were published in 2020 and then 2018 (83 and 80 documents, respectively). Only 61 were published in 2019. Production of the three document types showed fluctuations during the study period. As for conference papers, two peaks exited: one in 2016 and one in 2019; 72 papers were published in each (Figure 11). In general, their production rate is higher than that of articles. As for the reviews, there was a constant low level of production. As can be seen, 2020 was the year with the highest level of production (four documents).

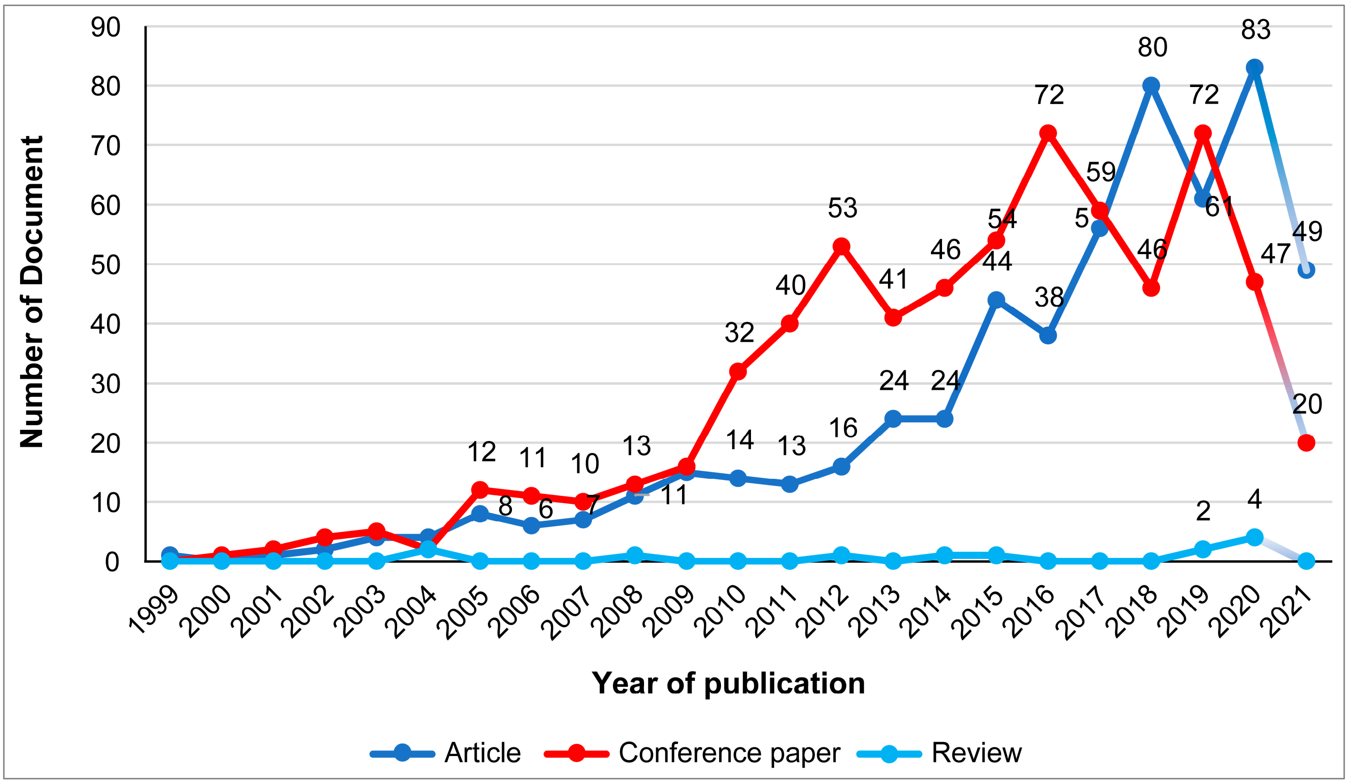
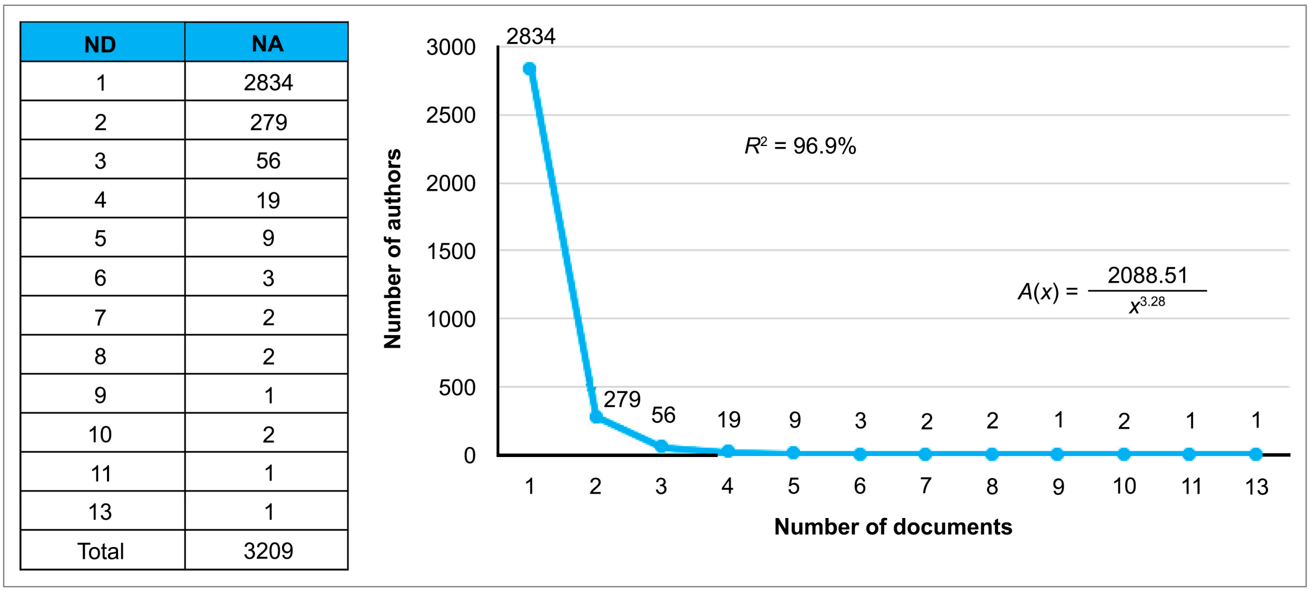


Figure 11. Evolution of scientific production (1999–2021) according to documentary typology.

5.1.2. Most Productive Authors

As can be seen in Figure 12, among the total 3209 authors of the 1231 papers analyzed (Figure 12), 2834 published only one paper, and only one author published 13. This shows that there is a large number of authors with low productivity and a small group of authors who publish a larger number of papers on competencies and professional skills in the area of BI&A in organizations. Additionally, the inverse Lotka model was used, having a goodness-of-fit index (coefficient of determination) of r^2 = 96.9%, which indicates that the model fits the analyzed data.

As a complementary analysis, the most productive authors (Table 5) of the community examined (seven or more documents) were identified and compared by their main institutions of affiliation and countries, in addition to their Scopus IDs and h-indexes. Most authors are affiliated with institutions in Spain, Canada, France, Austria, Italy, the United States and China. The h-index represents an indicator that uses two indicators (production and number of citations) to show the performance of a researcher based on the distribution of citations in their articles published over a period of time. The h-index values shown in the table include the total production of each author in journals indexed in Scopus. Thus, authors such as John Mylopoulos (53) and Yong Shi (43) have higher h-indexes than the others, which reflects the higher visibility of their papers published in journals indexed in Scopus.



ND	NA
1	2834
2	279
3	56
4	19
5	9
6	3
7	2
8	2
9	1
10	2
11	1
13	1
Total	3209

Figure 12. Productivity of authors according to number of published papers using Lotka's inverse model. *Note*: ND: number of documents, NA: number of authors.

Table 5. Top 9 most productive authors by number of papers, institution, country, Scopus ID and h-index.

N°	Author	NT	Main Institution & Country	Scopus ID	H-Index *
1	Trujillo, Juan Carlos	13	Universidad de Alicante Spain	7103051196	29
2	Mylopoulos, John	11	University of Toronto Canada	7005652259	53
3	Bimonte, Sandro	10	Université Clermont Auvergne France	15074087900	14
4	Maté, Alejandro	10	Universidad de Alicante Spain	42961909600	14
5	Schrefl, Michael	9	Johannes Kepler University Linz Austria	6603818133	17
6	Carta, Salvatore Mario	8	Università degli Studi di Cagliari Italy	7004254388	24
7	Saia, Roberto	8	Università degli Studi di Cagliari Italy	56029094200	14
8	Goul, Michael	7	W. P. Carey School of Business United States	6701579478	16
9	Shi, Yong	7	Chinese Academy of Sciences China	7404963015	43

Note: NT: number of texts on the subject, * Scopus h-index, 10 February 2022.

5.1.3. Journals with the Highest Scientific Output

For this analysis, only scientific journals were considered. Of 384 total journals, 11 were the most productive (Table 6), each having a minimum of 5 publications on the subject in the study period. *Decision Support Systems* and the *International Journal of Information Management* are the most productive journals with 12 publications each, and they are ranked in quartile 1 (Q1) SJR 2020 of Scopus. As for country, the United Kingdom and the United States have the most productive journals. Among publishing houses, Elsevier has the highest frequency of publications, having three journals in quartile 1.

Table 6. Most productive journals by country, number of documents, quartile and publisher.

No.	Journal	Country	NT	Quartile *	Publisher
1	*Decision Support Systems*	Netherlands	12	Q1	Elsevier
2	*International Journal of Information Management*	United Kingdom	12	Q1	Elsevier
3	*Expert Systems with Applications*	United Kingdom	10	Q1	Elsevier
4	*Communications of the Association for Information Systems*	United States	8	Q2	Association for Information Systems
5	*IEEE Access*	United States	7	Q1	Institute of Electrical and Electronics Engineers
6	*Journal of Intelligence Studies in Business*	Sweden	7	Q2	Halmstad University
7	*Journal of Computer Information Systems*	United Kingdom	6	Q1	Taylor and Francis
8	*Sustainability (Switzerland)*	Switzerland	6	Q1	MDPI AG
9	*Journal of Database Management*	United States	5	Q3	IGI Publishing
10	*Management Decision*	United Kingdom	5	Q1	Emerald Group Publishing
11	*Information Professional*	Spain	5	Q1	The Information Professional

Metacharacter *: Represents ranges of values 1 to 4.

5.1.4. Indicators of Collaboration

This analysis expresses the dynamics of the evolution of scientific collaboration in terms of the number of authors per document. Of the 1230 papers, 13.1% were written by a single author, 26.4% were written by two authors, 29.1% were written by three authors and 31.4% were published by four or more authors (Table 7).

Table 7. Distribution of publications by year and number of authors.

No. of Authors	Year of Publication								Total
	1999–2001	2002–2004	2005–2007	2008–2010	2011–2013	2014–2016	2017–2019	2020–2021	
1	0	3	9	15	29	38	48	19	161 (13.1%)
2	1	5	12	38	54	70	95	50	325 (26.4%)
3	1	9	19	24	51	88	115	51	358 (29.1%)
≥4	3	6	14	24	54	84	118	83	386 (31.4%)
Total	5	23	54	101	188	280	376	203	1230

To further analyze the behavior of collaboration according to publication periods, Figure 13 shows the three indicators that describe this behavior: the collaboration index (CI), the degree of collaboration (DC) and the collaboration coefficient (CC). The top graph shows the CI values indicating the average number of authors per document. In the 1999–2001 period, the highest values were recorded, the average being 3.1 authors per document. From the 2008–2010 period to the 2020–2021 period, there was slight linear growth in the number of authors per paper. The lower part of Figure 13 shows the DC values ranging from 0 to 1. In all periods, more than 80% of the documents were written collaboratively and globally. In turn, 87% of documents were published in collaboration. The CC combines the benefits of the two previous indicators, in addition to considering the difference between authors. The overall value was 0.57, and its maximum value was 0.72 in the 1999–2001 period. In general, there is a growing trend toward collaboration in this field of research.

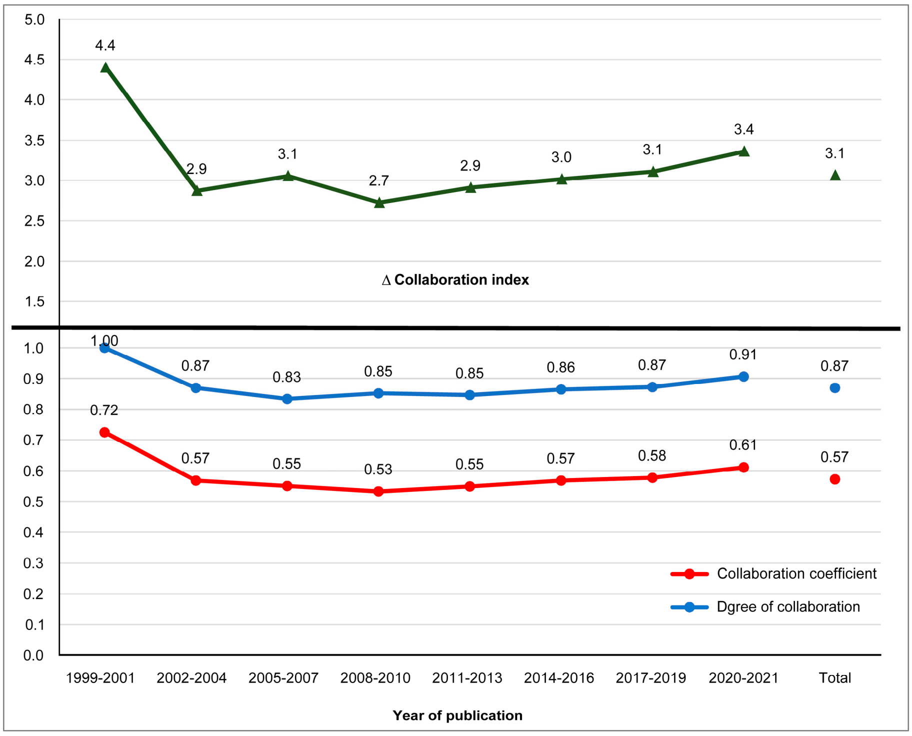


Figure 13. Collaboration index (CI), degree of collaboration (DC) and collaboration coefficient (CC) of authors (1999–2021).

5.1.5. Visualization of the Network and Keyword Overlay

Co-occurrence analysis was performed on 182 keywords from a population of 2539, which met the threshold of at least three occurrences. In the analyzed documents, each node represents a keyword, and its size is proportional to its number of occurrences. The top part of Figure 14 shows the keyword network visualization map. The colors indicate clusters of related keywords, according to the association strength method provided by the VOSviewer program. A total of 10 clusters, differentiated by colors, were obtained. The terms *business intelligence*, *data mining*, *data warehouse*, *big data* and *OLAP* had the highest frequency of occurrence as the central themes of the study, with 589, 145, 135, 118 and 80 occurrences, respectively.

As can be seen in Figure 14, the business intelligence node is the one with the highest number of links. In other words, its centrality is given by the number of links and the position it occupies in the network. In this direction, words such as data warehouse, data mining, big data, olap, decision support system (dss) and machine learning appear. In effect, the centrality (high degree) shows that the studies of the last 20 years have been occupied with relevant research on issues related to BI&A from the technical dimension and with little interest in relating them to management issues. In fact, if we look at Figure 14, words related to management are the ones with the lowest number of links and are positioned at the margins of the network. Balance scorecard, monitoring, reporting, data management, integration and motivation are some examples of words that describe the situation.

The overlay visualization of Figure 15 shows the use of keywords in the documents based on the annual average of the publication that covers the years 2012 to 2018. The evolution of the most important themes according to color can be seen. Thus, the blue color depicts—on average—the most frequent terms in the documents published since 2012 until reaching the red color in 2018. The primary terms include knowledge management (2012); data warehouse, OLAP (2013); data mining, conceptual model and e-commerce (2014); database, decision support system and association rules (2015); data analytics, analytics, business analytics and cloud computing (2016); clustering, big data, big data analytics and sentiment analysis (2017); and ML and data science (2018).

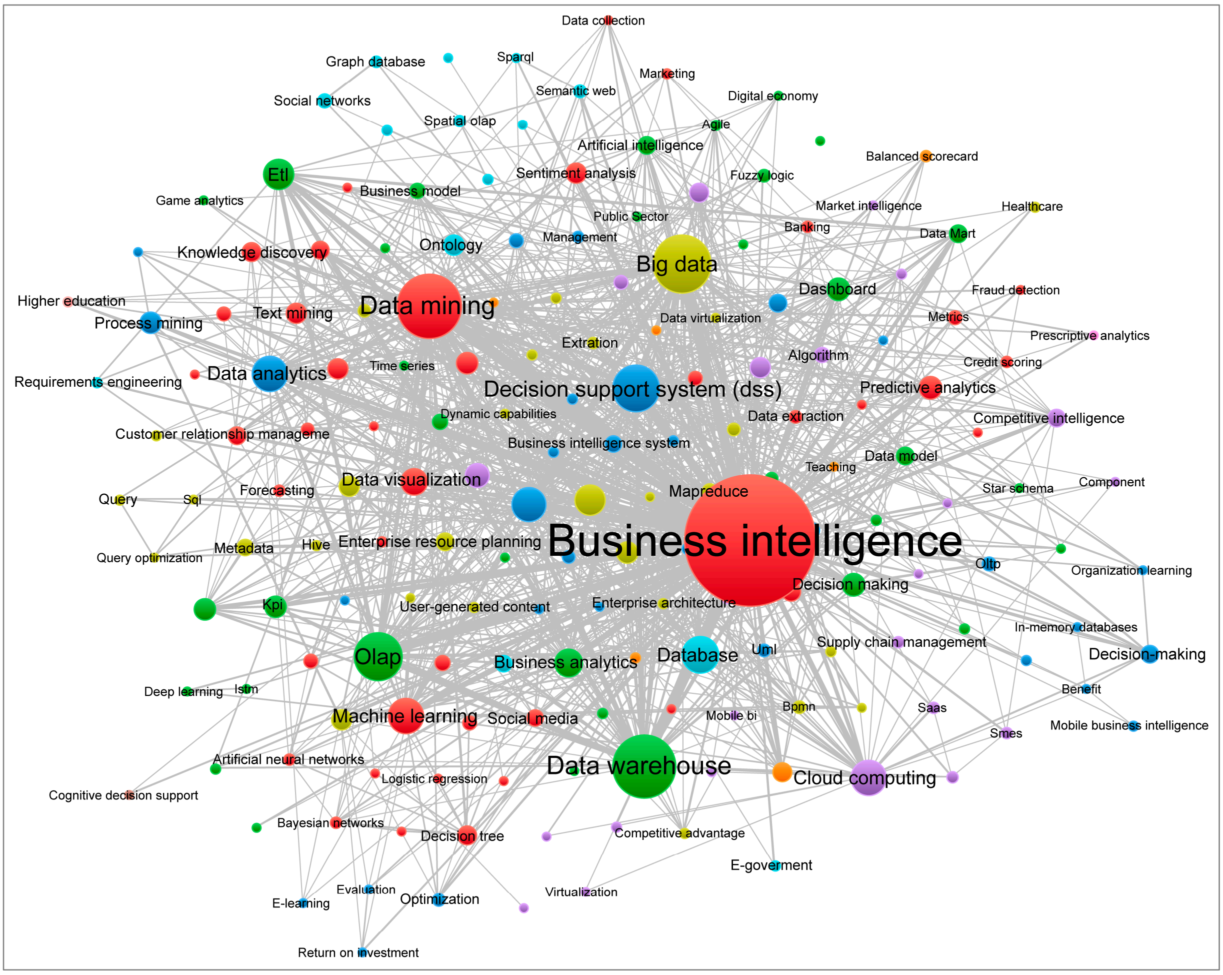


Figure 14. Network visualization map.

Similarly, as shown in Figure 14, the most frequent topics, both established and emerging, continue to be focused on words associated with BI&A from a technical perspective. There is little presence of words linking BI&A to management.

5.2. Results of Stage 2

5.2.1. Recognition of Main Study Topics in Scientific Production from 1999 to 2021: LDA Model

As can be seen in Figure 16, a first group emerged with three topics that showed greater and growing interest (topics 7, 12 and 16). Indeed, the pattern of topic 7 showed sustained growth since the period 2006–2008 and a steep upward slope during the period 2015–2017. Topic 7 was present in all 20 years of scientific production review. The terms reflected in this topic are related to the domain of industrial applications and empirical studies that are linked to organizational concepts such as capabilities, enterprise and performance (Figure 6).

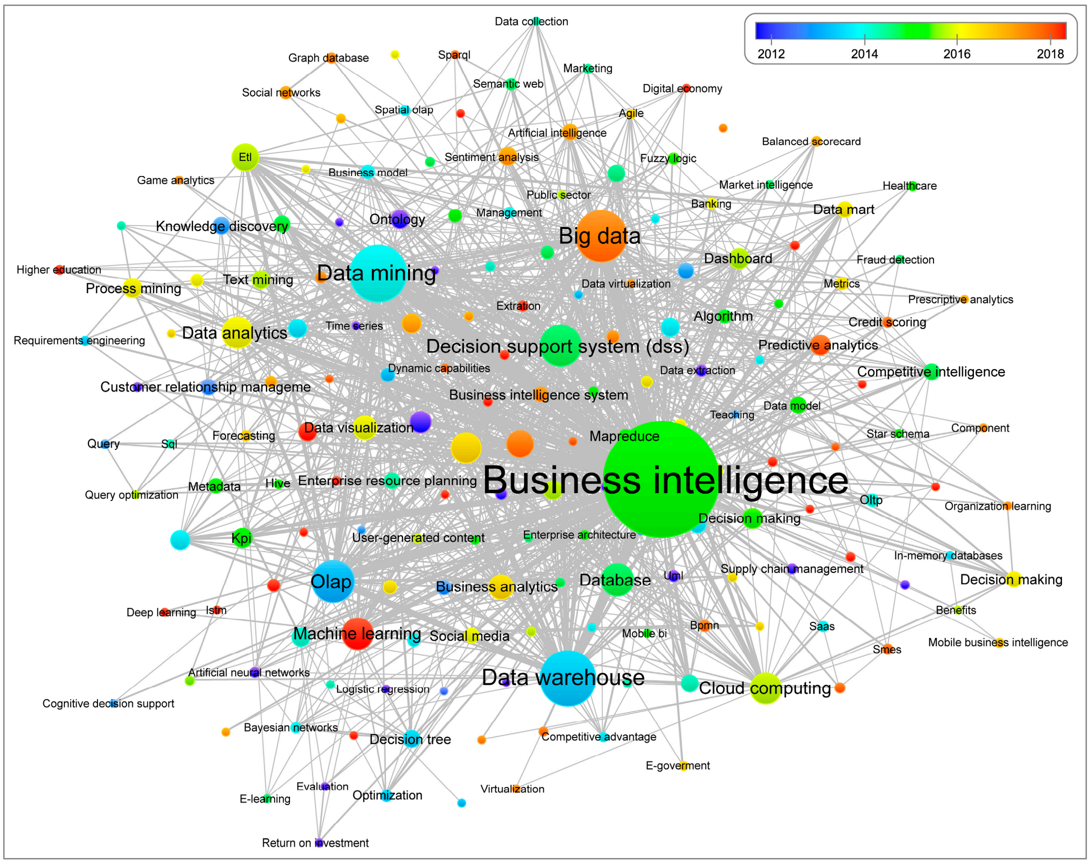


Figure 15. Keyword overlay map.

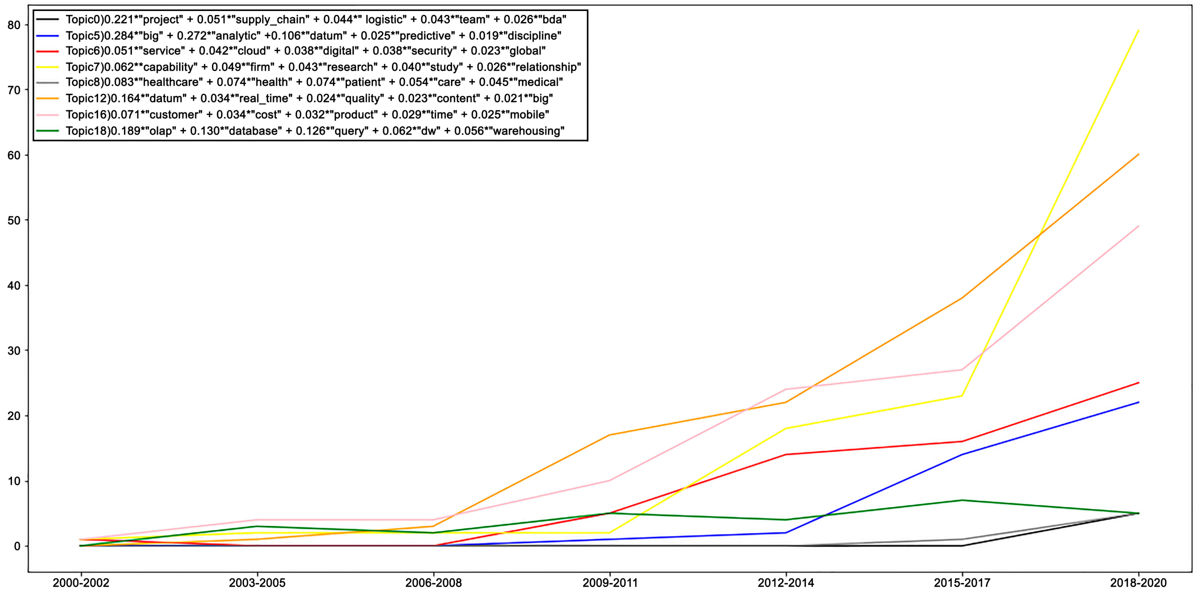


Figure 16. Graph of the pattern of behavior of the main topics discovered.

The second topic showing a high level of interest in Figure 16 is topic 12, for which there was a clear increase in the number of papers; highlighted terms include data, content, processing, real time and unstructured (see Figure 6). Figure 16 shows that since the 2006–2014 period, the topic has experienced considerable growth and has been particularly relevant from 2014 till today.

Finally, in Figure 16, topic 16 shows greater, albeit moderate, growth compared to the others since the 2000–2006 period. Subsequently, since 2006–2008, it had a sustained upward trend; it was the most important during the 2012–2014 period. This topic is centered on studies focused on the company, showing high frequencies of the words customer, mobility, products and costs (Figure 6).

Next, Figure 16 shows a second group (topics 6 and 5) with two topics of growing interest, albeit more moderate than the previous one. In fact, in topic 6, a fairly moderate pattern of interest can be observed during the 2000–2011 decade, and from then onward, there has been sustained growth. This topic is focused on IT architecture/infrastructure for BI, mainly cloud computing, services, security, digital transformation and security (see Figure 6). Figure 16 shows that topic 5 established its presence during the 2006–2008 period; there was an increasing trend from the 2012–2014 period through when the study was conducted. This topic includes the terms analytical approach and big data (Figure 6).

The third group (topics 18, 8 and 0) can be seen in Figure 16. Within the group, topic 18 had interest since the 2000s, although less so in the two groups mentioned above and with a decreasing trend in the last five years. The terms of the highest presence, the terms were queries, OLAP, data warehouse and storage (see Figure 6). Meanwhile, topic 8 showed little significant interest from 2000 to 2014. Since 2014, although there has been slight growth in recent years, this could be explained by COVID-19. The main terms in this vector are sanitation, health, patients and pandemic criteria (Figure 6).

Finally, topic 0, as can be seen in Figure 16, presents a similar pattern to the previous one, probably because of the same reason as topic 8. It is associated with supply chain business processes and logistics issues (Figure 6).

5.2.2. Main Clusters of Scientific Production, Articles from 1999 to 2021: Text Clustering

Figure 17 presents clusters (clusters 4, 1 and 2), and a growth pattern of significant interest can be observed. The strongest trend is of cluster 4, which is focused on data modeling and analytics. Then, there is cluster 1 for BI in general and cluster 2 associated with big data.

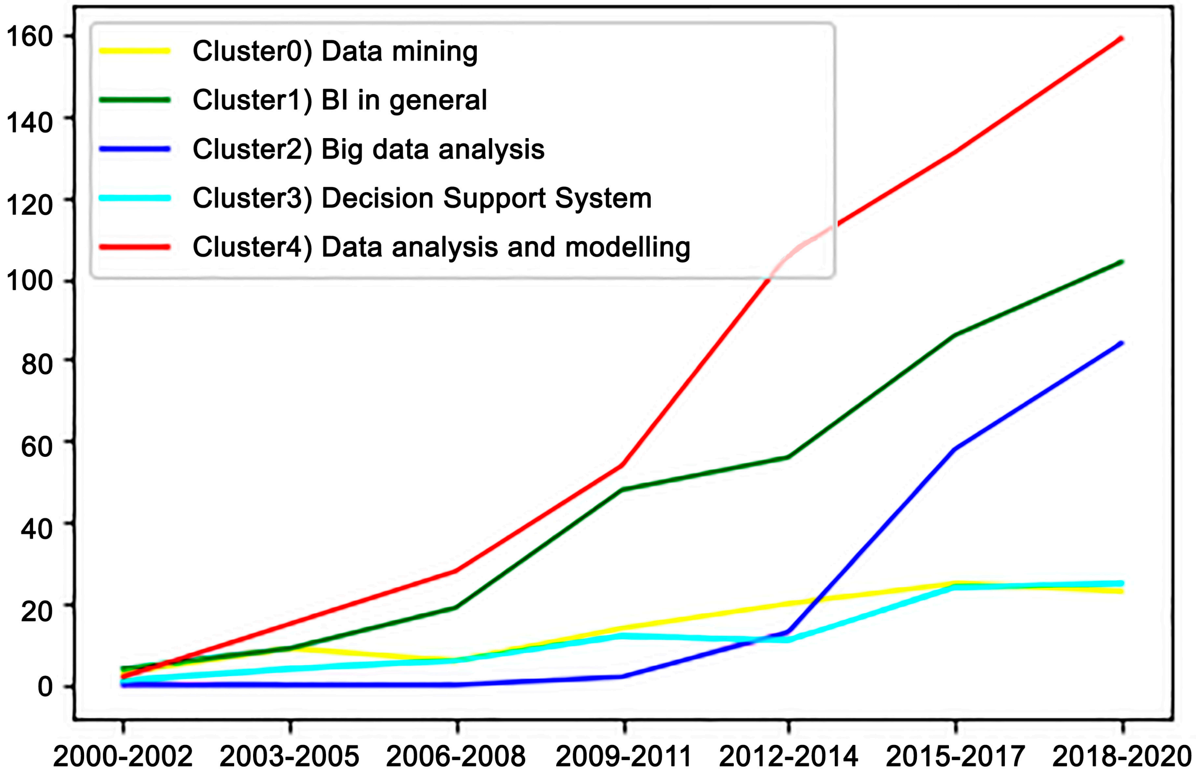


Figure 17. Graph of the distribution of all the clusters (five article groups) over time.

Clusters 0 and 3 can also be observed. Cluster 0, which corresponds to data mining, shows moderately increasing interest from 2000 to 2015, and remains steady will the current date. In turn, cluster 3 is associated with decision support systems, showing moderate variations in the pattern of behavior with respect to interest between 2000–2015 and then remaining stable.

6. Discussion

We aimed to shed light on the scientific production carried out over the last 20 years in BI&A associated with the concepts of strategic leadership competencies, with the objective of understanding the development of this field of research and all those skills necessary for professionals who exercise organizational management with BI&A in order to create competitive advantages during the decision-making process.

For this reason, a metric analysis was developed, in which a search protocol was applied and information was retrieved from the world's most recognized multidisciplinary database platforms, Web of Science (WoS) Core Collection and Scopus, without a time restriction.

The analysis corresponding to the informetric methodological stage (Stage 1) focused on metrics associated with scientific activity, i.e., scientific collaboration, structural analysis and thematic structure. Indeed, the results of scientific activity according to document typology show that the three document typologies analyzed were reviews, articles and conferences. The latter two have each shown an upward trend over the last 20 years of study, whereas reviews have shown a considerably decreasing trend over the same period. The production of conferences in general is higher than that of articles. In total, 561 are articles and 658 are conferences. Reviews in 2018 and 2020 showed increasing interest by constituting 80 and 83 papers, respectively.

The pattern of the productivity of authors by the number of published papers was analyzed under Lotka's inverse model [66], showing that 2,834 authors published only one paper; 279 authors published two and 13 papers were published by one author. With a goodness of fit r^2 = 96.9%, it is shown that there are many authors with low productivity and that a small group of authors publish most of the documents on competencies and professional skills in the field of knowledge in business intelligence (BA) and analytical intelligence (BI).

With respect to the measures of author collaboration in the study period, the number of authors per paper was 3.1 on average and showed an increasing trend from 2008–2010. The highest value was 4.4 CI for the period of 1999–2000. Additionally, more than 80% of the papers were written and published, showing a general trend towards collaboration.

The analysis of the machine learning results (Stage 2) from the application of the topic modelling algorithm (LDA) highlights three topics of increasing interest (topics 7, 12 and 16). In general terms, these three topics had empirical industrial domains with interest in large unstructured data processing with enterprise, market and customer orientations (Figure 6).

The second group (topics 6 and 5) is focused on predictive analytical techniques with enterprise, value and market orientations. These two groups in general terms show the interest in predictive analytical techniques, and in turn, in the need for infrastructure and technological architecture for enterprise transformation to support BI&A evolution.

The third group of topics (topics 18, 8 and 0) is associated with BI&A 1.0, in which it is observed that BI&A's interest to the areas of health and supply chain and logistics is increasing, which could be explained by COVID-19.

In the analysis of the results of the main clusters of the application of the k-means clustering algorithms, stage 2, of the scientific production in the last twenty years, a group that stands out was found; clusters 4, 1 and 2, which can be placed in a general way within the discovered topics (Figure 6), topics 5, 7, 12 and 16. The same behavioral pattern of scientific interest focused on empirical industrial domains and interest in unstructured big-data processing with orientations towards business, market and customer data analytics (Figure 18).

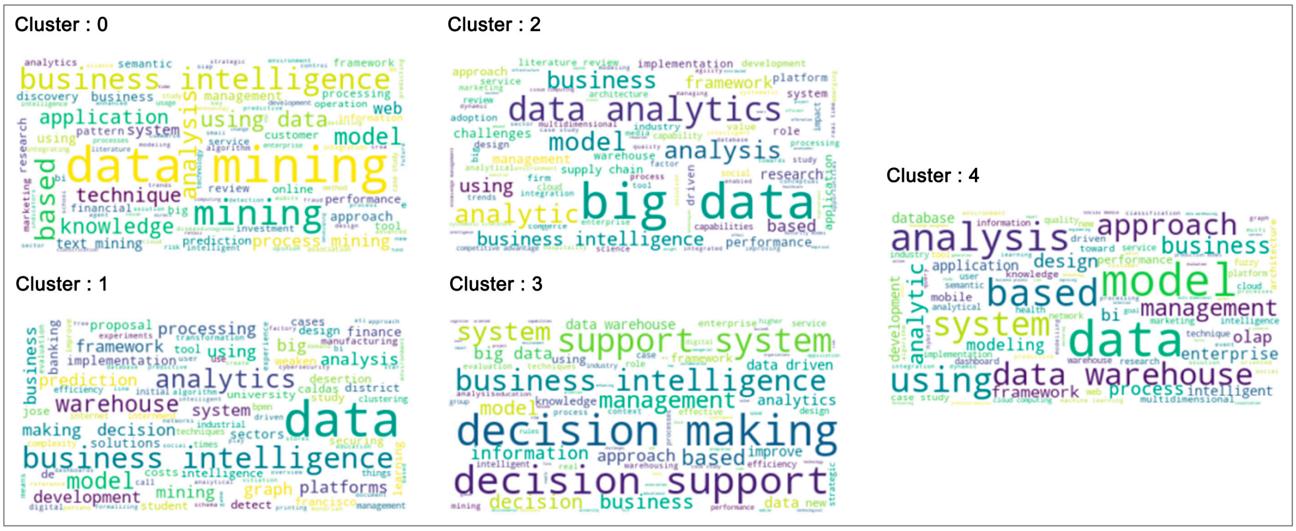


Figure 18. Predominant word cloud by discovered article cluster.

The results of the main clusters, after the application of k-means clustering (stage 2), with respect to scientific production over the last twenty years, showed a group that stands out: clusters 4, 1 and 2 can be located within topics 5, 7, 12 and 16. It shows the same behavioral pattern of scientific interest focused on empirical industrial domains, such as unstructured big-data processing with orientations towards business, market and customer data analytics (Figure 11).

Cluster 0 can be related to topics 18, 8 and 6 in Figure 6, which present interest in business, value and market-oriented predictive analytical techniques, and in turn, to the infrastructure and technology architecture needs for organizational transformation for BI&A. Finally, cluster 3 could be broadly associated with the themes discovered in topic 0, i.e., supply chain, logistics and projects supported by decision support systems (DSS) (Figure 11).

Look at Figure 16: a pattern of behavior can be seen in the eight main themes discovered in relation to the results in Table 7, with respect to the distribution of publications over 20 years. Indeed, there is a consistent relationship that shows that during the first 10 years there has been an increase in the number of studies in the field of strategic leadership competencies in BI&A. There has been a total of 101 publications (Table 7). In the same vein, the graph of topics in Figure 6 also shows a growing pattern of interest in topics such as big data, real time, data analytics, cloud, security, customers and mobility. Scientific production is predominant in this area: there have been 38 publications by two authors and 24 publications by three and four or more authors. In this sense, consistency can be observed when the increasing pattern of publications is present in two of the main groups of topics, modelling and data analysis and BI in general (Figure 17).

Now, at the end of the recent decade, in 2017–2019, as can be noticed in Table 7, the total number of publications in the field of study almost quadrupled, going from 101 publications to 376. Of them, a total of 143 publications were published by one or two authors and 233 publications by three or more authors. This increase happened from 2014 onwards (Table 7). In the same period, the results of topics obtained also showed strong growth and focuses on case studies, capacity, analytics, prediction and big data. Relating the topics to the results of the clusters confirmed the growing pattern of scientific production on competencies in BI&A belonging to the study clusters on big data, data modelling and analysis and BI in general (Figure 17). The most productive journals in research on strategic leadership competencies in BI&A were to be found based in The Netherlands, the United Kingdom, the United States, Sweden and Spain (Table 6).

The results of the keyword network analysis of the scientific production (Figure 14) show that the most recurrent groups of central themes are business intelligence, data mining, data warehouse, big data and olap, and show consistency with the results obtained from the application of the unsupervised algorithmic technique (LDA) (Figure 6), which picked out analytical capacity, prediction and big data. Likewise, the thematic evolution having the highest presence between 2012 and 2018 (Figure 15), when applying the informetric techniques, resulted in the following themes: database, decision support system and association rules (2015); data analytics, analytics, business analytics and cloud computing (2016); clustering, big data, big data analytics, predictive analytics and case study (2017); and machine learning. Indeed, these themes are entirely consistent with the findings from the results of the text clustering algorithm, confirming the pattern of clusters of study themes into big data, data modelling and analytics and BI in general (Figure 17).

7. Conclusions

In summary, with regard to the results of the descriptive study (stage 1) of the scientific production on strategic leadership competencies in BI&A in the last 20 years, it can be affirmed with an r^2 = 96.9% that a small group of authors published most articles. Of the total number of authors, the majority produced little on the subject of study.

With regard to the results of our study applying machine learning (ML) and natural language processing (NLP) techniques, eight topics were identified with good scores, which were focused on empirical industrial domains focusing on unstructured big-data processing oriented towards the company, market and customer. The other topics discovered (5 and 6) can be said to be oriented towards predictive analytical techniques and studies of a company's technological infrastructure and architecture to support the evolution of BI&A, and these topics (8 and 0) very likely emerged due to the COVID-19 changes. On the one hand, BI&A techniques relate to health issues, and on the other hand, to logistics and distribution.

Taking as a reference the LDM-BI&A model [40] that specifies leadership and BI&A competencies composed of six dimensions—(I) professional capability development domain for business managers in BI&A; (II) pro-business environment domain for business managers' learning about BI&A; (III) integrating BI&A skills for their expert work (mental habits) domain; (IV) BI&A strategic vision domain—a relationship can be established by classifying the key result terms present in the semantic network (Figure 14) with each of the competency dimensions of the LDM-BI&A model, under a general criterion on competencies, given that the key terms are not written in a context of action to put into practice. A proportion of the key terms present in the semantic network is obtained for each dimension of LDM-BI&A [22].

Consequently, it is possible to conclude that for the domain enabling an environment for business managers to learn about BI&A, related to establishing an adequate working environment that allows peer support, identification, dissemination of good practices and active learning of technologies associated with business intelligence and analytics or BI&A, the proportion of key terms present in the semantic network is low, at 0.72%. Similarly, the proportion of key terms considered as competencies associated with the domain that integrates BI&A skills into expert work (habits of mind), which is related to analytics for solving organizational problems, is considerably poorly represented, at only 3.01%. Finally, the BI&A strategic vision domain represents 3.37% of the keywords. It linked to the ability to think creatively about the future and to analyzing the effects of external (political, social, cultural and economic aspects of the country) and internal (of the organization) factors or variables.

The results show a high proportion for the domain of the development of professional capabilities for business managers in BI&A: 92.29 %. This shows that the development of professional capabilities is the most represented dimension, representing a set of technical and theoretical skills and knowledge on business intelligence and analytics, data analysis, statistical techniques (descriptive, bivariate, inferential, etc.), platforms and applications associated with business intelligence (BI).

Therefore, this study sheds light on the state of the art of BI&A in the field of strategic competencies for organizational leadership, contributing to informing readers that the dimensions of strategic leadership in BI&A is a fertile and largely unexplored field. This study can be used as a framework to advance the design of new research.

8. Limitations and Suggestions

We recommend advancing research involving concepts and theories underlying other disciplines focused on leadership competencies in BI&A, with a focus on professional competencies for business managers in BI&A. This could help progress our understanding of the development of competencies to enable an adequate work environment involving peer support, identification, dissemination of good practices and active learning of technologies associated with BI&A. It could also assist in further understanding the development of competencies of leaders to establish an enabling environment for learning in BI&A and integrating BI&A skills in expert work, thereby organizing BI&A for the sake of a strategic vision.

It is recommended to enhance this research methodology with future studies on the analysis of scientific production involving the application of advanced AI techniques and models, and to develop multidisciplinary research that integrates leadership, human resources and management theories with BI&A studies. In that direction, the work of Musarra, et al. (2022), regarding emotions, cultural intelligence and mutual trust in technology and business relationships, concludes that partner companies that freely express and understand each other's emotions and feelings are more likely to have a relationship characterized by mutual trust and confidence, and that each party will deliver on its obligations and promises. In other words, a range of leadership skills have a significant effect on business performance, particularly in the context of the major development that BI&A is experiencing [67].

Consequently, extending and integrating various theoretical bodies and methodologies for a thorough understanding of the reciprocal effects of leadership and BI&A on business value creation is a challenge that needs to be addressed in future research.

A limitation of this work is the need to increase the validation by experts so that they can analyze and classify the terms present in the keyword network in order to place them in a specific (not general) context in each dimension of the LDM-BI&A model.

Author Contributions: Data curation, M.O.F. and H.d.l.F.-M.; formal analysis M.O.F. and H.d.l.F.-M.; investigation, M.O.F. and H.d.l.F.-M.; methodology, M.O.F. and H.d.l.F.-M.; writing—original draft, M.O.F. and H.d.l.F.-M.; writing—review and editing, M.O.F. and H.d.l.F.-M. All authors have read and agreed to the published version of the manuscript.

Funding: Hanns de la Fuente-Mella was supported by a grant from the Núcleo de Investigacionen Data Analytics/VRIEA/PUCV/039.432/2020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

References

1. Côrte-Real, N.; Ruivo, P.; Oliveira, T. The diffusion stages of business intelligence & analytics (BI&A): A systematic mapping study. *Procedia Technol.* **2014**, *16*, 172–179. [CrossRef]
2. Wixom, B.; Watson, H. The BI-based organization. *Int. J. Bus. Intell. Res.* **2010**, *1*, 13–28. [CrossRef]
3. Chen, H.; Chiang, R.H.L.; Storey, V.C. Business intelligence and analytics: From big data to big impact. *MIS Q.* **2012**, *36*, 1165. [CrossRef]
4. Olszak, C.M. Toward better understanding and use of business intelligence in organizations. *Inf. Syst. Manag.* **2016**, *33*, 105–123. [CrossRef]
5. Davenport, T.H. From analytics to artificial intelligence. *J. Bus. Anal.* **2018**, *1*, 73–80. [CrossRef]
6. Burgess, A.J. *The Executive Guide to Artificial Intelligence: How to Identify and Implement Applications for AI in Your Organization*; Palgrave Macmillan: Cham, Switzerland, 2018.
7. Grover, V.; Chiang, R.H.L.; Liang, T.-P.; Zhang, D. Creating strategic business value from big data analytics: A research framework. *J. Manag. Inf. Syst.* **2018**, *35*, 388–423. [CrossRef]
8. Davenport, T.H. Competing on analytics. *Harv. Bus. Rev.* **2006**, *84*, 98.
9. Olszak, C.M.; Ziemba, E. Critical success factors for implementing business intelligence systems in small and medium enterprises on the example of upper Silesia, Poland. *Interdiscip. J. Inf. Knowl. Manag.* **2012**, *7*, 129–150. [CrossRef]
10. Dinh, J.E.; Lord, R.G.; Gardner, W.L.; Meuser, J.D.; Liden, R.C.; Hu, J. Leadership theory and research in the new millennium: Current theoretical trends and changing perspectives. *Leadersh. Q.* **2014**, *25*, 36–62. [CrossRef]
11. Lussier, R.N.; Achua, C.F. *Leadership: Theory, Application & Skill Development*, 6th ed.; Cengage Learning: Boston, MA, USA, 2016.
12. Yammarino, F. Leadership: Past, present, and future. *J. Leadersh. Organ. Stud.* **2013**, *20*, 149–155. [CrossRef]
13. Bolden, R.; Gosling, J.; Marturano, A.; Dennison, P. *A Review of Leadership Theory and Competency Framework*; Centre for Leadership Studies, University of Exeter: Exeter, UK, 2003.
14. Northouse, P.G. *Leadership: Theory and Practice*, 6th ed.; SAGE Publications: Thousand Oaks, CA, USA, 2012.
15. Paulienė, R. Interaction between managerial competencies and leadership in business organisations. *Reg. Form. Dev. Stud.* **2021**, *21*, 97–107. [CrossRef]
16. Boyatzis, R.E. Beyond Competence: The choice to be a leader. *Hum. Resour. Manag. Rev.* **1993**, *3*, 1–14. [CrossRef]
17. McClell, S. Gaining competitive advantage through strategic management development (SMD). *J. Manag. Dev.* **1994**, *13*, 4–13. [CrossRef]
18. Spencer, L.M.; Spencer, S.M. *Competence at Work: Models for Superior Performance*; John Wiley & Sons: Nashville, TN, USA, 1993.
19. Black, S.A. Qualities of effective leadership in higher education. *Open J. Leadersh.* **2015**, *04*, 54–66. [CrossRef]
20. Bennett, N.; Lemoine, G.J. What a difference a word makes: Understanding threats to performance in a VUCA world. *Bus. Horiz.* **2014**, *57*, 311–317. [CrossRef]
21. Dondi, M.; Klier, J.; Panier, F.; Schubert, J. Defining the Skills Citizens Will Need in the Future World of Work. McKinsey Global Institute. Available online: https://www.mckinsey.com/industries/public-and-social-sector/our-insights/defining-the-skills-citizens-will-need-in-the-future-world-of-work (accessed on 30 June 2021).
22. Faúndez, M.O.; de la Fuente-Mella, H. Skills Measurement Strategic Leadership Based on Knowledge Analytics Management through the Design of an Instrument for Business Managers of Chilean Companies. *Sustainability* **2022**, *14*, 9299. [CrossRef]
23. Wang, Y. Business Intelligence and Analytics Education: Hermeneutic Literature Review and Future Directions in IS Education. 2015. Available online: https://papers.ssrn.com/abstract=2603365 (accessed on 19 November 2022).
24. Ardito, L.; Scuotto, V.; Del Giudice, M.; Petruzzelli, A.M. A bibliometric analysis of research on Big Data analytics for business and management. *Manag. Decis.* **2019**, *57*, 1993–2009. [CrossRef]
25. Di Vaio, A.; Hassan, R.; Alavoine, C. Data intelligence and analytics: A bibliometric analysis of human–Artificial intelligence in public sector decision-making effectiveness. *Technol. Forecast. Soc. Chang.* **2022**, *174*, 121201. [CrossRef]
26. Peifer, Y.; Jeske, T.; Hille, S. Artificial Intelligence and its Impact on Leaders and Leadership. *Procedia Comput. Sci.* **2022**, *200*, 1024–1030. [CrossRef]
27. Thomas, B.; Senith, S.; Alfred Kirubaraj, A.; Jino Ramson, S.R. Does management graduates' emotional intelligence competencies predict their work performance? Insights from Artificial Neural Network Study. *Mater. Today* **2022**, *58*, 466–472. [CrossRef]
28. Olszak, C.M. Business intelligence systems for innovative development of organizations. *Procedia Comput. Sci.* **2022**, *207*, 1754–1762. [CrossRef]
29. Nacke, O. Informatrie: Ein never name für eine disciplin. *Nachr. Dokum.* **1979**, *30*, 429–433.
30. Lotka, A.J. La distribución de frecuencias de la productividad científica. *Rev. Acad. Cienc. Wash.* **1926**, *16*, 317–323.
31. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
32. Aggarwal, C.C.; Zhai, C. Aggarwal, C.C.; Zhai, C. A survey of text clustering algorithms. In *Mining Text Data*; Springer: Boston, MA, USA, 2012; pp. 77–128.
33. Muller, A.E.; Ames, H.M.R.; Jardim, P.S.J.; Rose, C.J. Machine learning in systematic reviews: Comparing automated text clustering with Lingo3G and human researcher categorization in a rapid review. *Res. Synth. Methods* **2022**, *13*, 229–241. [CrossRef]

34. Carpineto, C.; Osiński, S.; Romano, G.; Weiss, D. A survey of Web clustering engines. *ACM Comput. Surv.* **2009**, *41*, 1–38. [CrossRef]
35. Stansfield, C.; Thomas, J.; Kavanagh, J. "Clustering" documents automatically to support scoping reviews of research: A case study: "Clustering" to support scoping reviews. *Res. Synth. Methods* **2013**, *4*, 230–241. [CrossRef]
36. Sarkar, D. Semantic Analysis. In *Text Analytics with Python*; Apress: Berkeley, CA, USA, 2019; pp. 519–566.
37. Aizawa, A. An information-theoretic perspective of tf–idf measures. *Inf. Process. Manag.* **2003**, *39*, 45–65. [CrossRef]
38. Bholowalia, P.; Kumar, A. Article: EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *Int. J. Comput. Appl.* **2014**, *105*, 17–24.
39. Aich, L.; Das, A. Informetrics of Webinars through Video Conferencing Platforms for Teaching and Learning by Different LIS Professional during COVID-19 Period: An Evaluative Study. *Libr. Philos. Pract.* **2021**, 6679.
40. Ebadi, A.; Auger, A.; Gauthier, Y. Detecting emerging technologies and their evolution using deep learning and weak signal analysis. *J. Informetr.* **2022**, *16*, 101344. [CrossRef]
41. Calof, J.; Søilen, K.S.; Klavans, R.; Abdulkader, B.; Moudni, I.E. Understanding the structure, characteristics, and future of collective intelligence using local and global bibliometric analyses. *Technol. Forecast. Soc. Chang.* **2022**, *178*, 121561. [CrossRef]
42. Żółtowski, D. Business intelligence in the balanced scorecard: Bibliometric analysis. *Procedia Inf.* **2022**, *207*, 4075–4086.
43. Zhang, J.Z.; Srivastava, P.R.; Sharma, D.; Eachempati, P. Big data analytics and machine learning: A retrospective overview and bibliometric analysis. *Expert Syst. Appl.* **2021**, *184*, 115561. [CrossRef]
44. Aboelmaged, M.; Mouakket, S. Influential models and deterministic models in big data analytics research: A bibliometric analysis. *Inf. Process. Manag.* **2020**, *57*, 102234. [CrossRef]
45. Mahesh, B. Learning Algorithms—A Review. *Int. J. Sci. Res.* **2020**, *9*, 381–386.
46. Ullah, I.; Liu, K.; Yamamoto, T.; Zahid, M.; Jamal, A. Machine learning modeling with SHAP approach for predicting electric vehicle charging station choice behavior. *Travel Behav. Soc.* **2023**, *31*, 78–92. [CrossRef]
47. Suominen, A.; Toivanen, H. Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *J. Assoc. Inf. Sci. Technol.* **2016**, *67*, 2464–2476. [CrossRef]
48. Chen, L.; Zhou, K.; Jing, J.; Fan, H.; Li, J. Solution path algorithm for twin multiclass support vector machines. *Expert Syst. Appl.* **2022**, *210*, 118361. [CrossRef]
49. Farkhod, A.; Abdusalomov, A.; Makhmudov, F.; Cho, Y.I. LDA-based topic modeling sentiment analysis using topic/document/sentence (TDS) model. *Appl. Sci.* **2021**, *11*, 11091. [CrossRef]
50. Mayilvahanan, K.S.; Takeuchi, K.; Takeuchi, E.; Marschilok, A.; West, A. Supervised learning of synthetic big data for li-ion battery degradation diagnosis. *Batter. Supercaps* **2022**, *5*, e202100166. [CrossRef]
51. Tseng, S.C.; Lu, Y.C.; Chakraborty, G.; Chen, L.S. Comparison of opinion analysis of review comments using unsupervised feature clustering via LSA and LDA. In Proceedings of the 2019 IEEE 10th International Conference on Awareness Science and Technology, Morioka, Japan, 23–25 October 2019.
52. Tiwari, S.; Agarwal, S. Life log data analysis using optimal feature selection-based unsupervised logistic regression (OFS-ULR) for chronic disease classification. *arXiv* **2022**, arXiv:2204.01281.
53. Montavon, G.; Kauffmann, J.; Samek, W.; Müller, K.-R. Explaining the predictions of unsupervised learning models. In *xxAI—Beyond Explainable AI*; Springer International Publishing: Cham, Switzerland, 2022; pp. 117–138.
54. Gupta, O.; Roos, G. Mergers and acquisitions through an intellectual capital perspective. *J. Intellect. Cap.* **2001**, *2*, 297–309. [CrossRef]
55. Lang, Z.; Liu, H.; Meng, N.; Wang, H.; Wang, H.; Kong, F. Mapping the knowledge domains of research on fire safety—An informetrics analysis. *Tunn. Undergr. Space Technol.* **2021**, *108*, 103676. [CrossRef]
56. Hood, W.W.; Wilson, C.S. The Literature of Bibliometrics, Scientometrics, and Informetrics. *Scientometrics* **2001**, *52*, 291–314. [CrossRef]
57. Ghahramani, Z. Unsupervised Learning. In *Advanced Lectures on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 72–112.
58. Ackley, D.H.; Hinton, G.E.; Sejnowski, T.J. A learning algorithm for Boltzmann machines. *Cogn. Sci.* **1985**, *9*, 147–169. [CrossRef]
59. Anderson, B.D.O.; Moore, J.B. *Optimal Filtering*; Prentice-Hall: London, UK, 1979.
60. Korenčić, D.; Ristov, S.; Repar, J.; Šnajder, J. A topic coverage approach to evaluation of topic models. *IEEE Access* **2021**, *9*, 123280–123312. [CrossRef]
61. Eck, N.J.V.; Waltman, L. Visualizing bibliometric networks. In *Measuring Scholarly Impact*; Springer: Cham, Switzerland, 2014; pp. 285–320.
62. Palacios Jimenez, P.H.; Mori-Diestra, K.E.; Limaymanta Alvarez, C.H.; Loyola Romaní, J.M.; Gregorio Chaviano, O. Análisis bibliométrico y de redes sociales de la Revista Peruana de Medicina Experimental y Salud Pública (2010–2019). *e-Cienc. Inf.* **2020**, *11*. [CrossRef]
63. Fujita, K.; Kajikawa, Y.; Mori, J.; Sakata, I. Detecting research fronts using different types of weighted citation networks. *J. Eng. Technol. Manag.* **2014**, *32*, 129–146. [CrossRef]

64. Hofmann, M.; Chisholm, A. (Eds.) *Text Mining and Visualization: Case Studies Using Open-Source Tools*; Apple Academic Press: Oakville, MO, USA, 2015.
65. Röder, M.; Both, A.; Hinneburg, A. Exploring the space of topic coherence measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining—WSDM '15, Shanghai, China, 2–6 February 2015; ACM Press: New York, NY, USA, 2015.
66. Nicholls, P.T. Bibliometric modeling processes and the empirical validity of Lotka's Law. *J. Am. Soc. Inf. Sci.* **1989**, *40*, 379–385. [CrossRef]
67. Musarra, G.; Kadile, V.; Zaefarian, G.; Oghazi, P.; Najafi-Tavani, Z. Emotions, culture intelligence, and mutual trust in technology business relationships. *Technol. Forecast. Soc. Chang.* **2022**, *181*, 121770. Available online: https://www.sciencedirect.com/science/article/pii/S0040162522002943 (accessed on 19 November 2022). [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.