mathematics

MDPI

Article

Epistemological Considerations of Text Mining: Implications for Systematic Literature Review

Daniel Caballero-Julia * and Philippe Campillo

ULR 7369—URePSSS—Unité de Recherche Pluridisciplinaire Sport Santé Société, Faculté des Sciences du Sport et de l'Éducation Physique, Univ. Lille, Univ. Littoral Côte d'Opale, Univ. Artois, F-59000 Lille, France; philippe.campillo@univ-lille.fr

* Correspondence: daniel.caballero-julia@univ-lille.fr

Abstract: In the era of big data, the capacity to produce textual documents is increasing day by day. Our ability to generate large amounts of information has impacted our lives at both the individual and societal levels. Science has not escaped this evolution either, and it is often difficult to quickly and reliably "stand on the shoulders of giants". Text mining is presented as a promising mathematical solution. However, it has not yet convinced qualitative analysts who are usually wary of mathematical calculation. For this reason, this article proposes to rethink the epistemological principles of text mining, by returning to the qualitative analysis of its meaning and structure. It presents alternatives, applicable to the process of constructing lexical matrices for the analysis of a complex textual corpus. At the same time, the need for new multivariate algorithms capable of integrating these principles is discussed. We take a practical example in the use of text mining, by means of Multivariate Analysis of Variance Biplot (MANOVA-Biplot) when carrying out a systematic review of the literature. The article will show the advantages and disadvantages of exploring and analyzing a large set of publications quickly and methodically.

Keywords: text mining; big data; systematic literature review; scopus; web of science

check for updates

Citation: Caballero-Julia, D.; Campillo, P. Epistemological Considerations of Text Mining: Implications for Systematic Literature Review. *Mathematics* **2021**, *9*, 1865. https://doi.org/10.3390/math9161865

Academic Editor: Bo-Hao Chen

Received: 15 June 2021
Accepted: 3 August 2021
Published: 6 August 2021

1. Introduction

The world of big data generates data in very large quantities very quickly and in all areas of knowledge. The main difficulty of big data being that it deals with interpreting an enormous amount of information of diverse origins. It is necessary to consider that a significant amount of time is correlated to this flow of data, to extract what is essential and useful [1]. This set, that is in perpetual quantitative expansion, progressively exceeds human processing capacities. Without the contribution of automatic algorithmic systems and other multivariate statistical mathematical models for processing and managing classic databases, the researcher, similar to any other individual (journalist, writer, librarian, etc.) who wishes to tackle any study will find themselves inundated with information and references.

In the initial task of building up a corpus of references centered on his problem, the researcher refers to classic bibliographic databases. They orient and condition "the state of the art" bibliography by their search engines, which work on the localization of terms mainly retained in the keywords, titles, summaries and other contents. Faced with this growing and massive production of textual documents, it becomes necessary to develop promising mathematical and software solutions which delve into quantitative sources to help extract the qualitative data [2].

Despite the obvious methodological interest, and a strong tendency to use text mining, there are, however, few historical and epistemological questions on qualitative and quantitative approaches in the analysis of textual corpora in databases. It is true that the application of information extraction to text is related to the problem of simplifying the text, in order to create a structured view of the information present in the free text. The problem of the quality of the information being retained, as well as the meaning being clear is fundamental, while the overall objective is to create a text more easily readable by the machines to process the sentences. The questioning initially raised on the quantitative/qualitative affiliation would make it possible to question the symmetry between qualitativism and quantitativism, in relation to the problem of meaning, as well as the nature of the linguistic knowledge produced, taking into account the context and the relevance of an approach in relation to the other on the processing of corpora.

It is necessary to remain vigilant and careful in the manipulation of data clusters, by interpreting the epistemological and ontological effects of big data. As Törnberg and Törnberg [3] pointed out, contemporary epistemological and philosophical discussions highlight, for example, the contradictions in the study of digital social life.

In literature reviews, especially "meta-analyses", authors refer to the mathematical tool by including "statistical meta-analyses", which facilitate the analysis of quantitative or qualitative data from different publications, in order to understand their variability. While meta-analysis was for a long time centered on analyses of an exclusively statistical nature, we can consider that qualitative meta-analyses have been deployed in other disciplinary fields since the 1990s.

According to the same logic of "knowledge extraction" in the bibliographic references, the methods are more specifically applied to the study of textual corpus (lexicometry, textometry, computational linguistics, text mining, etc.), and refer to the advances in mathematics in the functioning of artificial intelligence. These techniques are undergoing important developments in all scientific fields and in society in general. These methods are becoming increasingly important in the humanities and social sciences, particularly among researchers who claim to be part of the digital humanities.

For this reason, this article proposes to rethink, among other things, the methodological and in a certain way epistemological principles of text mining by returning to the qualitative analysis of its meaning and structure. It presents alternatives applicable to the process of building lexical matrices for the analysis of a complex textual corpus. At the same time, the need for new multivariate algorithms able to integrate these principles is discussed. We take a practical example in the use of text mining with MANOVA-Biplot [4–6] in a systematic literature review. This article will show the advantages and disadvantages of the fast and methodical exploration and analysis of a large number of publications.

2. The Text Mining (Textual Statistical Analysis)

The history of text mining involves research into methods, techniques and tools of linguistic analysis, in order to answer the problem of meaning to be extracted in the exploitation of often scattered and unstructured documents. It was in 1958 that the researcher Hans Peter Luhn [7] from the IBM company used data processing to automate a synthesis of texts and published a first article on this subject. He studied statistical measurement of the frequency of words and their distribution, to establish a relative assessment of the meaning of words and sentences. The most significant words, expressions and sentences, in terms of meaning, are extracted and make it possible to constitute a synthetic summary.

Today, many tools are being developed (gate.ac.uk, sas.com, alpha works.ibm.com, statsoft.fr) to help with information processing. They generate new topics of corpus processing; however, the notion remains relativity present in the final synthetic meaning, and especially, in the confidence of the veracity of the automatic work. Then, the epistemological problem arises, in the intriguing value of the quantitative in relation to the qualitative, or even the consistency of the meaning of the extraction of the qualitative into the quantitative [8]. Simply put, is a word or phrase repeated several times more valuable than others used only once? In this context of interpretation of a semiotic order of words and their meaning, the singular logic of expressions and metaphors then become pretexts for dizzying variations [9,10].

In this epistemological problem of meaning extracted from automatic information deduced from programmed algorithms, quantitative and qualitative methods differ not only by the nature of their knowledge but also by their object of analysis. Thus, according to a traditional qualitative logic, the analysis of a text involves the search for the meaning of the text, as well as its structure [11] (see Table 1).

Table 1. Quantitative and qualitative description. Source: Murillo & Mena [11].

Perspective	Distributive (Quantitative)	Structural (Qualitative)
Epistemology	Facts	Meaning units
Nature of knowledge	Analyse patterns and correlations	Analyse discourses and associated social meaning
Object of analysis	Study the distribution of phenomena Positivism	Studying the relationship between phenomena Constructivism
Method	Deductive	Inductive
Data collection techniques and instruments	Questionnaires, formal statistics, etc.	Focus groups, interviews, life stories . . .

Today, the reader-researcher, in his process of exploring large corpus to extract the essential, uses software utilities, allowing him to extract information (IE). It is a recent technology which seeks to meet a very old need: to acquire knowledge from often very voluminous texts. In this text mining process, methodological ambiguity is engaged by the association of two disciplines with distinct origins and histories, statistics and linguistics (semiotics).

The history of the evolution of methodological processes underlying the statistical analysis of textual data (ASDT) shows, among other things, a type of statistical analysis called “à la française”, initiated by the mathematician and statistician Jean-Paul Benzécri (1932–2019). This has become possible thanks to the advancement of new information processing technologies, and the growing power of computer tools [12].

This type of analysis notably uses statistical algorithms to analyze more or less complex texts from the frequencies of a generated matrix. The advantage of Benzécri’s proposal [13], the Correspondence Factor Analysis (CFA), was the expression by materializing the visualization of the proximities between the rows of the matrix, and the modalities of the variables placed on two (or more) factor axes. This statistical method, the main characteristics of which is to be multidimensional and descriptive, rationalizes the interpretation of texts by the researcher, thanks to the projection of the data cloud from a multidimensional space into a two-dimensional space, even if he does not know the subtleties of the method [12–15].

This tradition was carried out by Lebart, Berry, Morineau and Salem [16–18], who positioned the CFA at the height of a technique of choice, among a set of statistical methods most important for the textual data analysis. At the same time, the work of Max Reinert [19–22], a former student of Benzecri, takes up the CFA and the Ascending Hierarchical Classification (CAH) via the development of the ALCESTE software (Société IMAGE and Centre National de la Recherche Scientifique (CNRS), France), one of the IT solutions proposed to describe and deepen the laws of vocabulary distribution in texts. A few years later, Osuna [23], in connection with the work of Lebart, proposed the Biplot methods [24] as an alternative to Factorial Correspondence Analysis. This method also allows simultaneous representation of data in rows and columns, to represent an Xnxp matrix, projecting the data cloud from a multidimensional space into a two-dimensional space.

Today, there are many works that can be indexed around text mining, using this type of analysis to analyze a more or less consistent set of texts. For example, Dalud-Vincent [25] uses and critiques the Ascending Hierarchical Classification for the analysis of semi-structured interview data from sociological studies, using the ALCESTE software proposed by Reinert. Caballero-Julia et al. [26] propose the conversion of text from discussion groups into corrected frequency tables, and then analyzes them using the HJ-Biplot [27–29]; Martin et al. [30] more recently used other software, such as Sonal (Alex ALBER, Université F. Rabelais de Tours/Laboratoire CITERES (UMR CNRS) France) or TXM [31], for the analysis of this type of interview. Since 2009, with the initiative of Pierre Ratinaud [32], the same top-down hierarchical classification algorithms of Max Reinert have been used by the strong community of supporters of open access in the statistical system R, supported by the Python language since the software called IRaMuteq to generate multidimensional analyses of texts and questionnaires [33]. As a whole, the analytical methods employed in the work on the interpretation of textual corpora are generally inspired by Benzécri, and based on the counting of units to evaluate their importance, according to the number of repetitions.

The qualitative/quantitative alternative remains to be considered in text mining, to establish a principle of rigor in the collection and exploitation of their materials. It is undoubtedly a methodological pooling, requiring different skills that can be enriched, in particular, the processing of numerical series can constitute a substantial contribution to qualitative research. However, qualitative research adopts a particular relationship to theory because it aims to analytically generalize and not to statistically generalize by seeking to highlight mechanisms that can operate differently according to contexts and situations. It must make systematic use of counterfactual reasoning and plausible rival assumptions in theoretical analysis. On the one hand, by referring to quantitative data and variables, the qualitative approach can justify the choice of methodology, determine the unit of analysis itself, which must be defined according to the research question and justify the field of empirical investigation [34–36]. On the other hand, the expression of meaning in a text emerges between the choice of two major semantic units, the words but also the sentences, which structures speech between repetitions but also the meaning of a symbolic order, which gives value to the qualitative approach compared to the quantitative approach.

3. Big Data, Research and the Evolution of Science

The major phenomenon of the exponential increase in the information to be processed is revolutionizing the way we approach it, in order to constitute syntheses, as well as predict and anticipate scientific and social evolution. They can be summed up with the HACE theorem: the heterogeneity of a large volume of data, the autonomy of sources, the complexity and constant evolution of the relationships that exist between the data [37]. At the same time, this trend can be characterized by the five Vs: volume, velocity (speed), variety, veracity and value [38]. In this sense, humans are confronted not only with a large number of complex and diverse data, but also with a set of techniques and algorithms which allow the organization, exploration and exploitation of this data in a relatively short time [37].

Kune et al. [38] explain that the taxonomy of big data is made up of dimensions as large as they are varied. If we look at Figure 1, we see that this diversity affects not only the techniques or mathematical models used but also a large set of fields and scientific disciplines: IT, statistics, commerce, security, etc. In this universe of possibilities, science is no exception to this upward trend, noting a strong increase in the number of articles and scientific work produced in recent years (see Figure 2). Initially, one can think that this growing productive capacity will allow the development of science and, consequently, will have an impact on the multiplication of scientific productions, as well as on the publications which will be available to scientists. As a result, in a second step, the research will (if we have not already completed it) deal with the difficulties of exploration, management and exploitation of a growing number of articles which constitute its scientific background [39].

In this sense, the scientist will have to manage a large volume of works [40] which are produced faster, all the while increasing the variety of productions, for which it will be necessary to verify their veracity and their value. In this context of increasing complexity, the development of new methods of management and processing of the scientific corpus is positioned as a necessity to be approached and treated as quickly as possible [41]. Given that all scientific productions are presented in a textual format (articles, conference papers, manuals, books, etc.), new techniques and technologies for processing and textual analysis are emerging as an essential support in the field of research [42]. To this end, text analytics, one of the components of big data (see Figure 1, big data taxonomy), constitutes the process that makes it possible to derive information from textual sources [38].

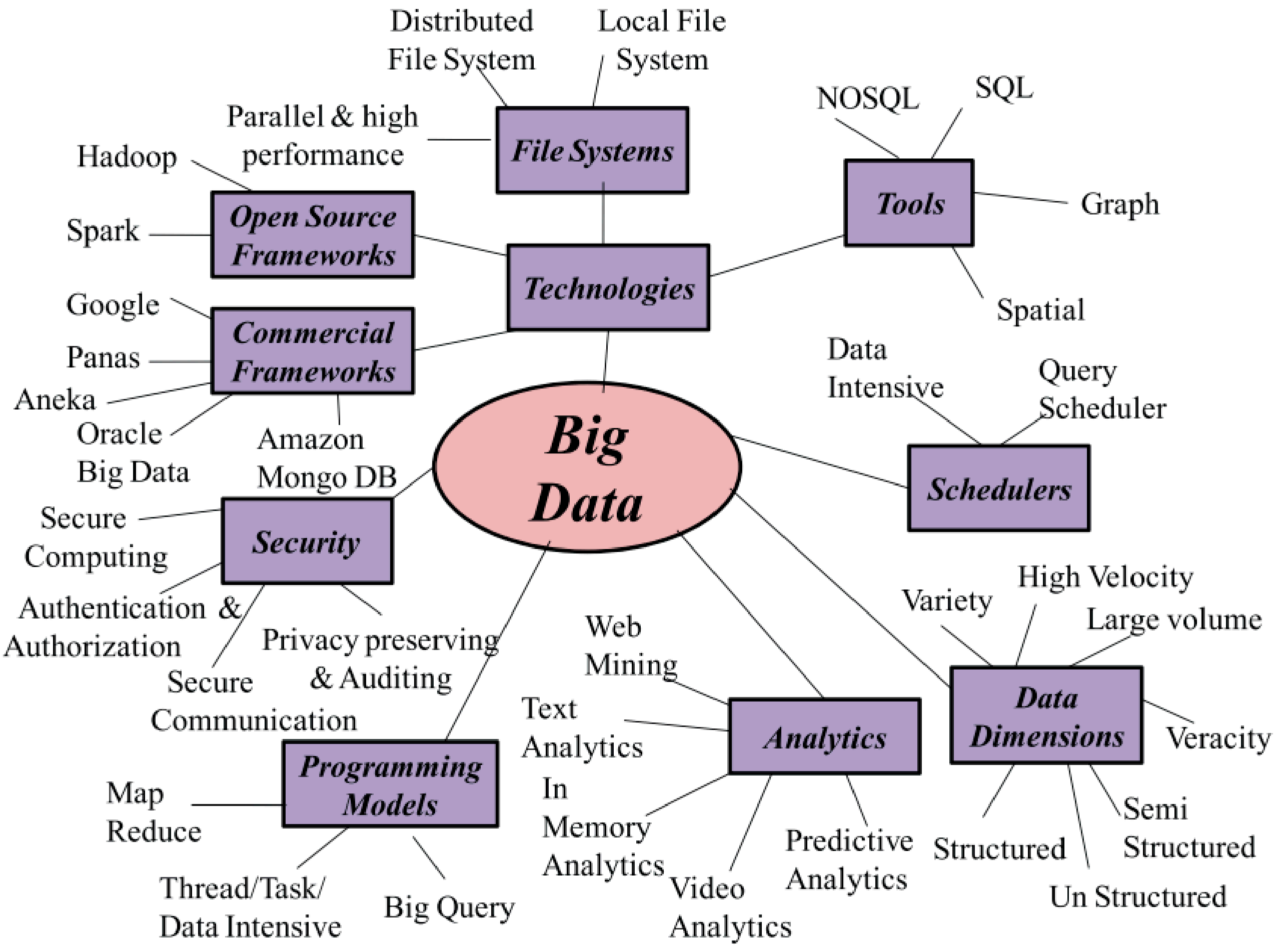

Figure 1. Big data taxonomy [38].

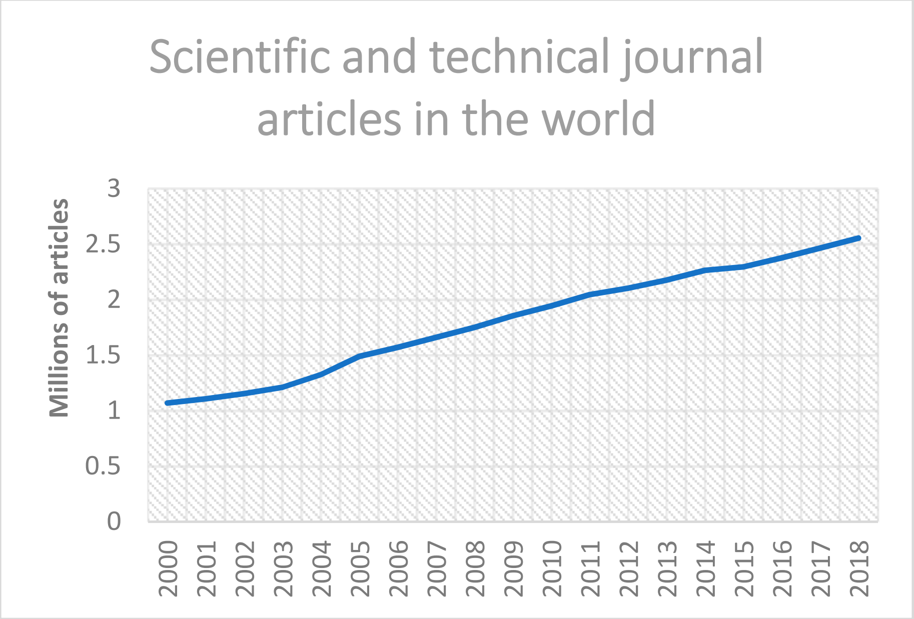

Figure 2. In the world. Source: the world bank, Ref. [40].

4. Systematic Literature Review (SLR)

As Snyder [39] asserts, the growing complexity of literature management tasks underscores the relevance of considering this literature review as a research method in itself. At the same time, the author insists on the need to systematize this review, in order to prevent scientists from working blindly, or based on principles that have already been refuted. As a result, it summarizes the principles, which are valid for all disciplines, which characterize the different types of literature reviews [43].

The development of an SLR proceeded according to the summary produced by Snyder [39], by formulating a set of questions that allow scientists to ensure the systematization of the stages of the review. In this sense, as a first step, we must conceive of research by asking ourselves about its contributions and the interest it arouses. At the same time, it is important to clearly define the search method and strategies (search terms, databases, inclusion and exclusion criteria, etc.). In a second step, the researcher will have to establish a practical and self-critical research plan, which will make it possible to establish an action plan and evaluate the selection process. Then, in a third step, we can strive to define the analysis criteria of our SLR. Finally, in a fourth step, our actions must lead us to clear and sufficient communication of our results, according to the objectives and the target audience determined in the first part.

In addition, as recalled by Munn et al. [43] citing the Cochrane manual, “a systematic review uses explicit, systematic methods that are selected with a view to minimizing bias, thus providing more reliable findings from which conclusions can be drawn and decisions made” [44]. In this sense, the need not only to systematize literature reviews but also to develop tools and techniques capable of facilitating and systematizing processes becomes a relevant and urgent subject, especially in the era of big data. This article therefore aims to propose one of these possible solutions that will allow scientists to appropriately explore complex, variable, changing and diverse literature. This solution uses the foundations of text mining to systematize and simplify the literature review.

5. Methodology

In this article, we propose a new solution for multivariate statistical textual analysis which can, among other things, be used to carry out a systematic review of the literature. In its design, we tried to integrate the epistemological considerations that we made in text mining, in order to provide a technique capable of generating knowledge while minimizing information processing times as well as biases. The reflexive but also methodological nature of this work justifies the broadening of the methodological explanations that develop in the different stages, as we approach the operations necessary to achieve success.

5.1. Step 1: Design of the Systematic Literature Review

As indicated in the previous part, we must start from the design of our research, by reflecting on the relevance and interest of the subject, the target audience to which we are directing our literature review and the criteria for selecting sources and works. In this article, we show two examples from two different research themes associated with the same project, and two additional examples from two different areas, in order to illustrate the possibilities of the technique and highlight its prospects for improvement and evolution.

The first two examples used here are part of a project focusing on the analysis of the representations and behaviors of actors (individuals and sports organizations), in cooperation and conflict relations that take place daily in the strategic development of women’s sport. The aim of the project is to discuss the challenges that women and girls still face in sport, in order to consider the future of women’s sport. Indeed, in a field historically built on exclusion, women’s sport has become institutionalized from the confrontation with discourses oscillating between the prescription and the proscription of the setting in motion of the female body. In this sense, the management of sports organizations evolve in a spatiotemporal context in constant transformation. The analysis of the impact of the representations of sports organizations on their decisions and repercussions of their strategies in the development of women's sport has therefore become a major subject in the discipline of sports management. For the design of this project, we needed to target two key aspects of the management of women's sport: that of professionalization and that of inequalities in sport. In this sense, these examples target an audience of sports managers and analyze the impact in the literature available to scientists.

In both cases, the articles were retrieved from the two most important databases internationally, where the journals are indexed, namely: Web of Sciences (WOS) and SCOPUS. These databases are considered to be juggernauts when it comes to indexing databases [45], and each currently includes more than 20,000 active journals [46–48].

Articles were selected using search engines integrated into the Web of Science and Scopus websites. As inclusion criteria, we proposed the same keywords in English in both databases, by selecting the widest possible period, while keeping the same exclusion criteria. As a result, we were able to retrieve 1283 articles which corresponded to the keywords of professionalization, 384 to those of inequality, 6053 to those of text mining. In both cases, the results were checked for relevance by excluding articles that did not match the topics (see Tables 2 and 3).

Table 2. Systematic literature review design I.

Specifications for SRL	Professionalization of Women's Sport	Inequalities in Women's Sport
Database	Web of Science Scopus	Web of Science Scopus
Keywords	Sport; Women; Professional	Sport; Women; Inequality
Periods	All years to present	All years to present
Exclusion criteria	The physical aspect of a woman Woman's body The physiology of women The media coverage of sports Physical performance Women-specific workouts Sports nutrition Professions other than sports Incomplete bibliographic information (abstract, year of publication).	Inequalities in sport outside women's sport. Gender inequalities outside sport. Women's sport excluding inequality. Incomplete bibliographic information (abstract, year of publication).
Number of articles obtained	1283 (1257 WoS; 26 Scopus)	384 (179 WoS; 205 Scopus)
Number of articles retained	128 (115 WoS; 13 Scopus)	165 (70 WoS; 94 Scopus)

Table 3. Systematic literature review design II.

Specifications for SRL	Text Mining	Low Back Pain Prevention
Database	Web of Science	Web of Science Scopus Pubmed
Keywords	Text Mining	Low Back Pain Education Prevention
Periods	The last five years (2017–2021)	All years to present
Exclusion criteria	All the articles not using or not describing any text mining technique Not an article document Articles without an abstract or year of publication	Medical treatment of pain Unrelated articles Articles without an abstract or year of publication
Number of articles obtained	6053 references where 3690 for articles	141
Number of articles retained	3521	77

To compare with other areas and illustrate the possibilities of this method, we introduce two additional topics. One from the health sciences, low back pain prevention, called "the illness of the century", a relevant topic in the prevention of illness. And a second (multidisciplinary) subject that is text mining, which is the subject of this article.

In this article, we focus on a simple systematic review of the literature, although we can do a more specific or complex one. So, for every example, we ask for the evolution of the literature and the differences between each period. To do this, we use the year of publication to make groups, in order to ask:

1. Does every single publication talk about a different issue, or are there general topics that appear in the publications?
2. Can we simplify the analysis of the literature?
   a. Are there vocabularies that are specific to each period?
   b. What are they talking about?
   c. Can we identify some articles that appear in every period? Which ones?

5.2. Step 2: Developing a Research Plan

From the results obtained using the search engines of the WoS, Scopus and PubMed websites, the site servers were asked to download a bibliographic matrix having, at a minimum, the summary of the articles, the year of publication, the name of the authors as well as of the journal. The researcher will be able to define a list of variables that interests them by personalizing the process. This matrix will make it possible to retrieve the data necessary for the analyses and the control of the selection process. For this, and in order to facilitate the management of the matrix, a new variable has been added, called ID (1, 2, 3, etc.), which allows each row of the matrix to be identified in a simple way.

5.3. Step 3: Definition of SLR Analysis Criteria

The construction of the lexical table is as follows.

Once we have retrieved and sorted the bibliographic matrix, we must begin to manage it using the principles of ASDT. This involves building a new data matrix from the original matrix. To do this, first of all, it is necessary to retrieve the summaries (text that will be analyzed in these examples) and the ID variable by creating a text file that is recognized by the IRamuteq software. IRaMuTeQ is a software which allows the use of the most common analyses of SDTA, and uses, among others, the CFA of Benzècri and the Reinert Method. This text document will take the following form:

**** *ID_1
Text of the first summary;
**** *ID_2
Text of the second summary;
**** *ID_3
Text of the third summary.

Where "****" indicates that we are going to introduce text, "* ID" that we will categorize this text using the variable ID and "_1" (2, 3, . . . , n) indicates the value that takes the variable at each time. This can be simplified by using a spreadsheet and using the formula: = "****" & "*" & "ID" & A2 & B2. Using A2, which corresponds to the value of the variable ID and B2 which corresponds to the value of the variable "summary". Once created, this document is fed into IRamuteq, which will apply a lemmatization protocol. This protocol puts all nouns and adjectives present in summaries in the masculine singular, and verbs in the infinitive. This makes it easier to understand and simplifies the resulting matrix. At the same time, the software removes words that are repeated only once, "hapax" as well as words considered as "empty words" (prepositions, conjunctions, etc.). This operation makes it possible to build an X_{mxp} frequency matrix with n words from the text and p articles, represented by the different IDs. This matrix is traditionally called a "lexical table" [16,17].

5.4. Data Characterization

As indicated in the previous parts, the algorithms proposed by the literature, such as CFA, CHD or Biplot Robuste, used this lexical table to initiate their analyses by favoring high frequencies. In this proposition, we use the characterization value described by Caballero et al. [26,49], which makes it possible to calculate the relative weights of each word in the matrix, according to their specific presence in an article instead of focusing only on high frequencies. This operation, which generates the recalculation of the original matrix, is perhaps the most important modification to the classical text mining process, given that it modifies the data weight to include the epistemological criteria (the search for the underlying meaning) in the calculations. It responds to the formula:

$$f'_{ij} = \frac{f_{ij}}{\sqrt{max_{f_i}}\sqrt{max_{f_j}}} \tag{1}$$

where f_{ij} represents the original frequency of the word located on row I and column j, and j and f'_{ij} is the new recalculated frequency. The original frequency is relativized to max_{f_i} et max_{f_j}, which represent the maximum of the frequencies of a row (words) and the maximum of the frequencies of a column (articles). The new X_{mxp} matrix benefits from the properties of this characterization value:

The new values are between zero and one.

If f_{ij} = max_{f_i} and = max_{f_j}

$$f'_{ij} = 1 \tag{2}$$

If $f_{ij} \neq max_{f_i}$ and = max_{f_j}

$$f'_{ij} = \sqrt{\frac{f_{ij}}{max_{f_i}}} \tag{3}$$

If f_{ij} = max_{f_i} and $\neq max_{f_j}$

$$f'_{ij} = \sqrt{\frac{f_{ij}}{max_{f_j}}} \tag{4}$$

If $f_{ij} \neq max_{f_i}$ and $\neq max_{f_j}$ f'_{ij} are equal to the characterization value.

In this way, being the highest frequency that can be found in this combination i j ensures the frequency of the maximum weight in the new matrix. However, being the maximum row or column without having both causes the new frequency f'_{ij} to be relativized to the other maximum i or j. Consequently, if a word does not represent any maximum, the characterization value calculates its new weight by relativizing it to the two maximums.

These properties make it possible to relativize the frequencies of the original matrix, according to their "singularity", instead of doing it starting from the high frequencies. At the same time, they make it possible to select the most characteristic words of the matrix by identifying the maximum values of each row. In this work, a minimum characterization of 0.7 was established as a selection criterion, in order to facilitate the reading of the graphical results and their analysis. Analysis of the lexical table is as follows.

The use of the characterization value is not compatible with CFA, Reinert method or any other technique that recalculates the weight of cells in a matrix. This would involve two recalculations on each original frequency. Among the large number of analyses which allow the exploitation of a lexical table, we sought one belonging to the family of biplots [24]. According to Osuna et al. [23,50], this family of analyses has advantages over, for example, CFA. The biplot algorithm uses the original matrix in these calculations, although CFA recalculates the weights of the original frequencies from the calculated row and column profiles.

$$f^i_j = k(i,j)/ki \text{ ou } ki = \sum_{j=1}^{n} k(i,j) \tag{5}$$

It does this for rows and symmetrically for columns, which prevents us from using CFA once we have used the characterization value that gives the matrix other weights. In addition, the biplot represents a two (three) dimensional plane from a Euclidean distance or a Mahalanobis distance, instead of using the chi-square distance between i and i' calculated by the CFA as

$$d^2(i,\ i') = \sum\left\{ \left(f^i_j - f^{i'}_j \right)^2 / f_j \middle|\ j \epsilon J \right\} \tag{6}$$

This facilitates reading by making it more intuitive. In addition, biplots, as a family, allow us to choose the technique that best suits the objectives of the research. This last characteristic seems to us a strong argument for selecting this type of analysis. In this case, the choice of MANOVA Biplot responds to the necessity to separate groups and see their differences by projecting data to the directions with the maximum power of discrimination between the groups.

In general, from the point of view of its computation, the biplot analysis described by Gabriel [24] is based on the singular value decomposition of a matrix $X_{n\times p}$ of rank r ($r \leq \min(n, p)$), where we may have the product

$$X_{(n\times p)} = U_{(n\times r)}\Lambda_{(r\times r)}{V'}_{(r\times p)}\text{with } U'U = V'V = I_r \tag{7}$$

where $U_{(n\times r)}$ is the matrix of eigenvectors of XX', $V_{(r\times p)}$ the matrix of eigenvectors of $X'X$ and $\Lambda_{(r\times r)}$ the diagonal matrix of $\lambda_1, \lambda_2, \ldots,\ \lambda_r$, of r eigenvalues of XX' or $X'X$. So that the elements of $X_{(n\times p)}$ are given by:

$$x_{ij} = \sum_{k=1}^{r} \sqrt{\lambda_k u_{ik} v_{jk}}\ i = 1,\ 2,\ \ldots,\ n,\ j = 1,\ 2,\ \ldots,\ p \tag{8}$$

Thus, the markers in the q dimension for the rows and columns of a matrix are:

$$J_q = U_q \Lambda_q \text{y } H_q = V_q \Lambda_q \tag{9}$$

For the analysis of our lexical table, we suggest the use of the MANOVA Biplot [5,51], also known under the name of Canonical Biplot [4,6]. This responds to the necessity to separate groups and see their differences by projecting data to the directions with the maximum power of discrimination between the groups. Certainly, this variant, "Allows a weighted representation of the matrix of means which has the directions with the maximum power of discrimination between the groups. This technique makes it possible to highlight the main differences between the groups, the variables responsible for the differences and the making of inferences under the canonical and original variables using circles of confidence located around the points representing the groups" [52].

The opposite case was used by Caballero [53], in order to analyze multiple tables and show the common structure with the help of the JK-Meta-Biplot [54].

The process takes place in three stages, starting from the data matrix which brings together all the information relating to different groups. In this case, the variable "year of publication" was used to serve as a guide for comparison between the articles published in the different periods. It was categorized into three groups (2000–2009 = 1; 2010–2014 = 2; 2015–2020 = 3) for the study of inequalities, two groups (before 2010 = 1; after 2010 = 2) for the study of professionalization, five groups for the study of text mining and six for the study of the low back pain. Therefore, we transposed the matrix and retrieved the variable selected from the ID assigned to each article. Thus, in a first step, we start from the matrix $X_g\left(I_g x\ J\right)$ generated from the lexical tables of Web of Science and Scopus, where each article is classified according to its year of publication (g), and we build the matrix of means $\overline{X}$ from the vectors of the means of g groups (g = 1, 2, . . . , g) with ng individuals (articles in our example) each (n = n_1, n_2, . . . , n_g). In a second step, we consider D the diagonal matrix, containing the size n of each group; the intra group covariance matrix is therefore expressed as:

$$W = \frac{1}{n-g}\left(X'X - \overline{X'} D_n \overline{X} \right) \tag{10}$$

and the intergroup covariance matrix as

$$B = \frac{1}{g-1}\overline{X'}D_n\overline{X} \tag{11}$$

And finally, in a third step, we consider the Singular Value Decomposition (SVS) as the matrix of the weighted means of the groups

$$Y = D_n^{1/2}\overline{X}W^{-1/2} \tag{12}$$

Resulting in the generalized form of the SVS of $\overline{X}$, where D_n et W^{-1} are used as metrics. This is how we can project the set of information onto a biplot representation

$$\begin{gathered}\overline{X} \cong GH' \\ \text{where } G = D_n^{-1/2}UD_\lambda \\ \text{and } H = W^{1/2}V\end{gathered} \tag{13}$$

Represents the row and column markers, respectively [4,5,51–56]. Moreover, as it has been defined by Amaro et at. [55] and recalled by Varas et al. [56], the coordinates of the centroid of group i defined in two dimensions by:

$$p_i^T = (P_{i,1}, P_{i,2})i = 1, 2, \ldots, g \tag{14}$$

We can construct a confidence region defined by a circle of radius $t_{n-g,\alpha}/\sqrt{n_g}$ where $t_{n-g,\alpha}$ represents the critical point of a Student's t with $n - g$ degrees of freedom and a confidence level of α.

These calculations are integrated into the MultBiplot v18.0312 software [54], developed by the University of Salamanca which was used to generate the one-way MANOVA Biplots that will be shown, by way of example, in this article. For this, we asked the software not to make additional modifications (raw data) in order to keep the matrix in the original state.

6. Results and Discussion

We intend to carry out a systematic review of the literature, as described in the methodology of this article. This means that we must analyze a large number of articles from scientific production relating to sport sciences (women's sport and more particularly to inequalities and professionalization in women's sport), health sciences (low back pain prevention) and a multidisciplinary topic (text mining). To do this, we finally collected nearly 8000 articles and processed a matrix for each subject generated from the text, resulting from the articles which correspond perfectly to the subjects discussed

6.1. Example 1: SRL in Sport Sciences (Inequalities and Professionalization of Women's Sport)

Our SRL of the inequalities in women's sport returned 165 articles (70 WoS; 94 Scopus), since 1980 to the present. According to the results of the classical methods, such as CFA (see Figure 3), many words are placed throughout the axes but the text from the abstracts was mixed, and it was difficult to read the content of the articles. The use of Reinert's method allowed us to identify two profiles: on one hand, we can see the words "experience", "field", "relation"," challenge", "social", "explore", "femininity", "identity", "masculinity", among others; on the other hand, we find the words "factor", "analyze", "south", "participation", "federation", "female", "find", "equality", "country", "level", "result", "Asian", among others. In addition, it was not easy to analyze the evolution of the literature and the differences between each period, which was one of our objectives.

However, the analysis of the matrix created from the lexical table on inequalities in women's sport revealed that there are significant differences between the three predefined groups from the years of publication. We can see those differences by looking at the distances between the means of the groups and the circles of confidence located around the points representing the groups. Thus, in Figure 4, we can observe that these three groups are clearly separated, which results in a distinctive vocabulary specific to each period. In other words, the themes that have been approached historically are not the same, and can be inferred from the graphic representation of the MANOVA Biplot.

For the interpretation of these themes, a graphic representation without vectors was generated (see Figure 5), in order to facilitate its reading and understanding. Thanks to this, we realized that the publications produced before 2009 focused on analyses of inequalities from a structural and systemic point of view.

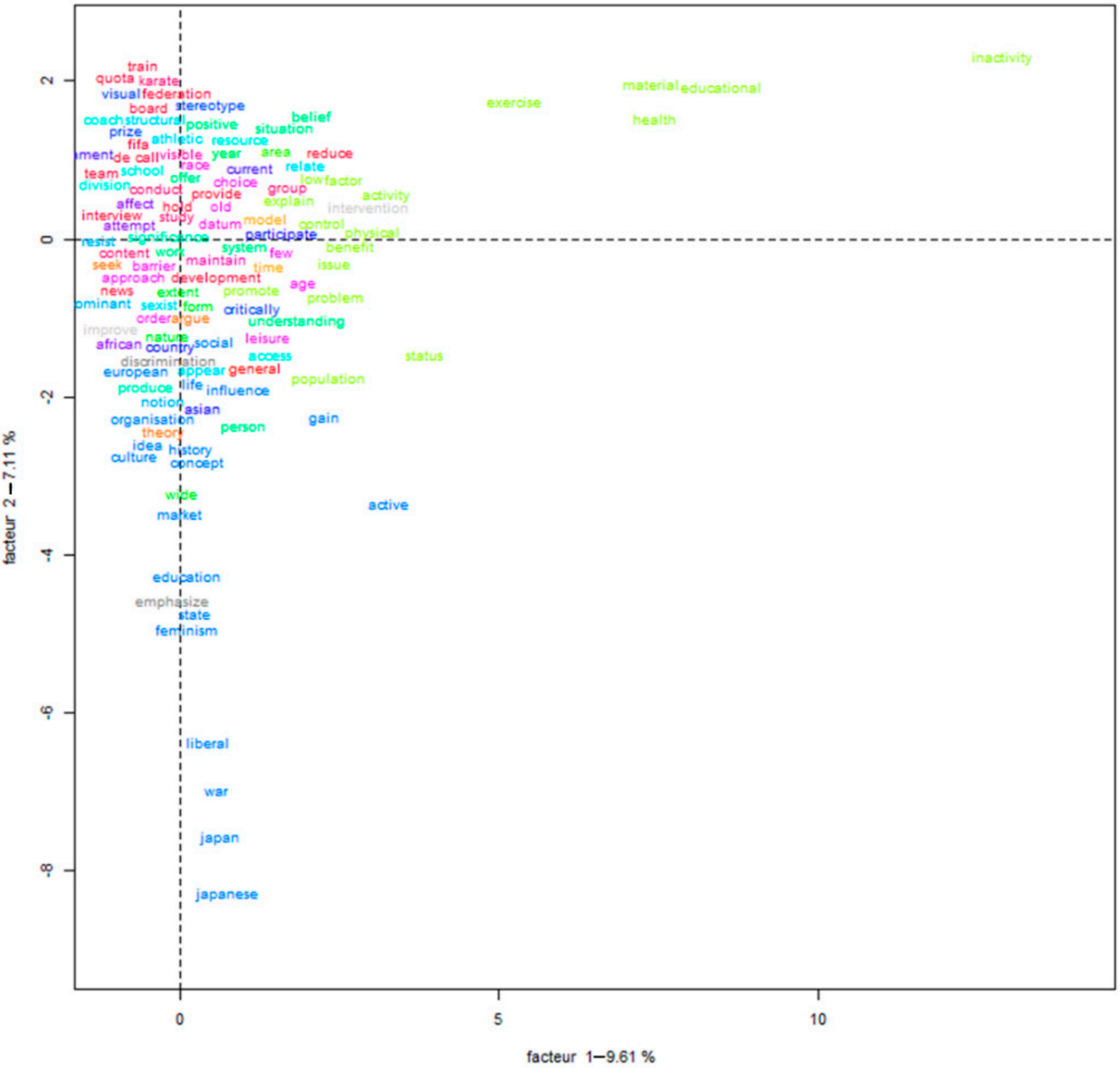

Figure 3. SRL of the inequalities in women's sport using CFA.

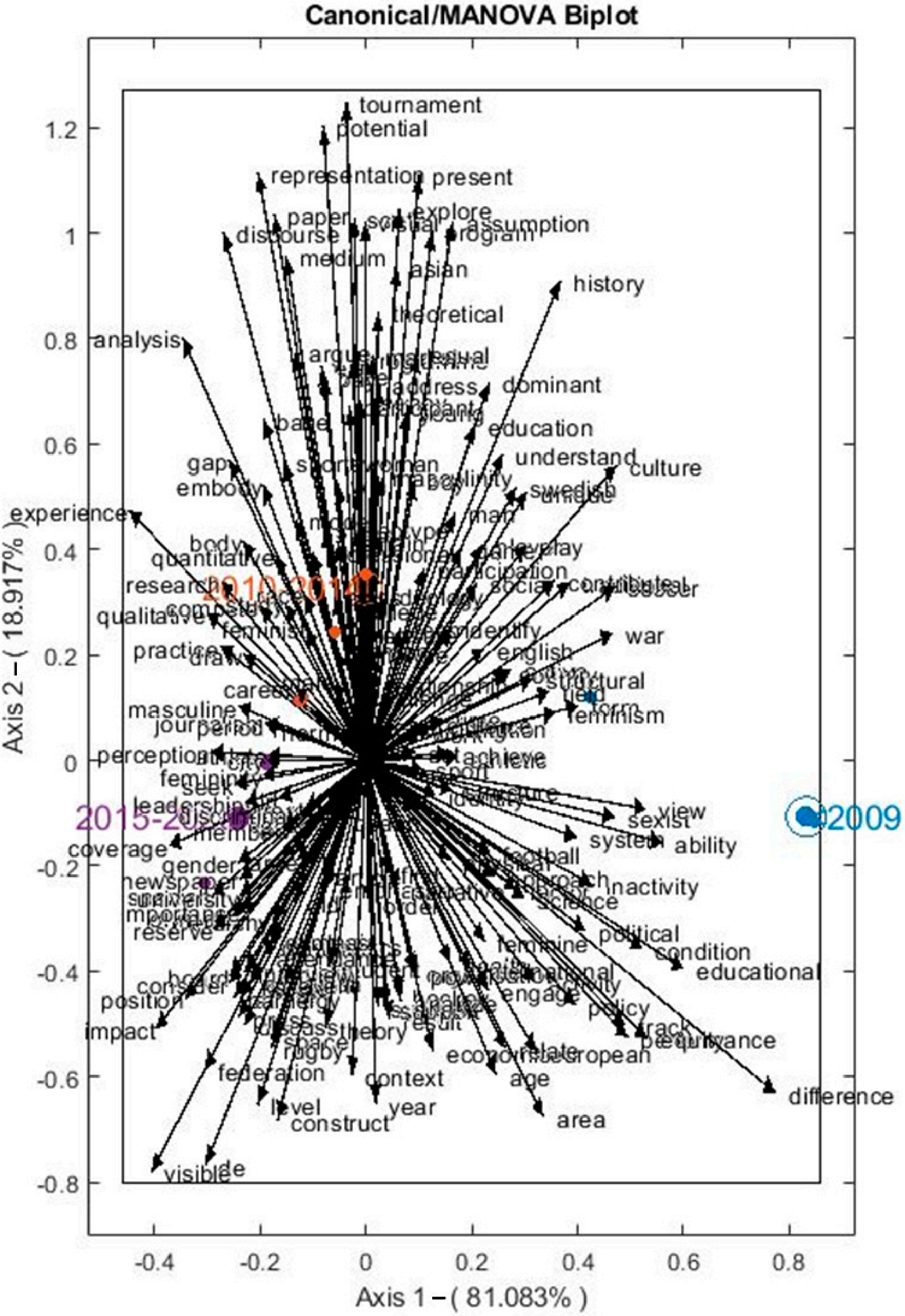

Figure 4. Manova Biplot for SLR inequalities in women's sport.

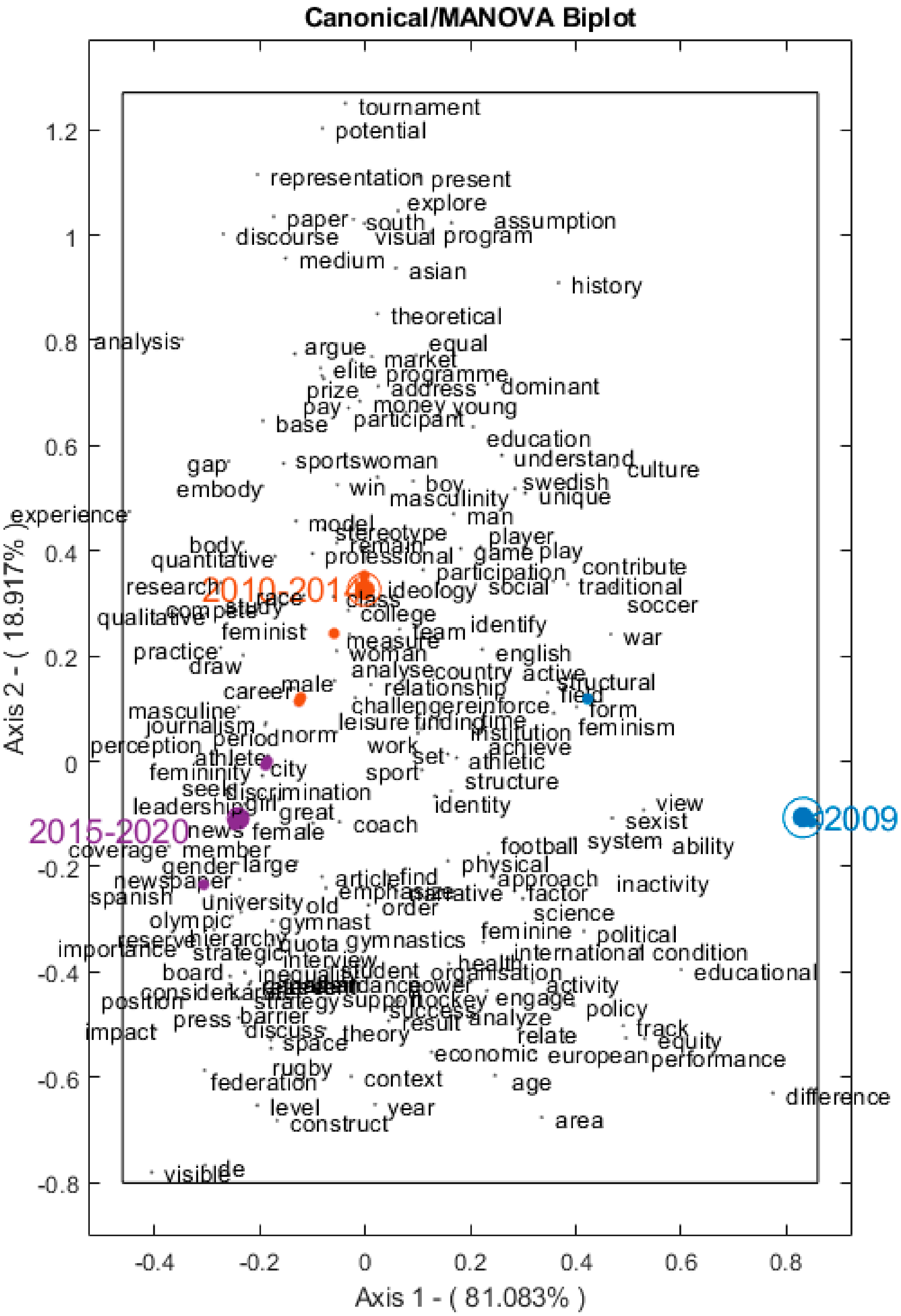

Figure 5. Manova Biplot without vectors for the SLR of inequalities in women's sport.

The words which characterize this group and which point towards the fourth quadrant are, among others: "view", "sexist", "system", "ability", "inactivity", "political", "condition", "educational", "difference". From 2010, the interest of researchers in inequalities in women's sport will turn to the analyses of the dominant representations and discourses in sport, a theme that can be reconstructed from words that have strong positive correlations with the positive values of the second canonical variable of the MANOVA Biplot (first and second quadrants). These words are, for example, turning, potential, representation, present, discourse, dominant. In this sense, we can consider that the literature dedicated to the study of women's sport seems to have shifted from a macro-analytical perspective to a more micro-analytical one, where qualitative and quantitative studies are more focused on discourse and representations. More recently, since 2015, the characteristic vocabulary of the articles selected is clearly differentiated from the other two, by aligning with the first axis but by opposing the group of publications released before 2009. This is a sub-corpus of publications on much more targeted aspects of gender inequalities in the media. We can clearly distinguish between the knotting of the words, "newspaper", "news", "journalism", "coverage", "impact", "position", "importance", among other words associated with feminism and gender studies such as "gender", "masculine", "femininity", "female", "discrimination".

As we can see in Figure 6, the scientific articles on the inequalities in women's sport have evolved from a macro sociological perspective (before 2009) of the inequalities to a micro sociological perspective (2010–2014), just before shifting to an analysis of inequalities focused on the analysis of the women's sport communication (2015–2020). The MANOVA Biplot shows us many articles (represented by little colored points) that we can also consider as transition articles between the main periods.

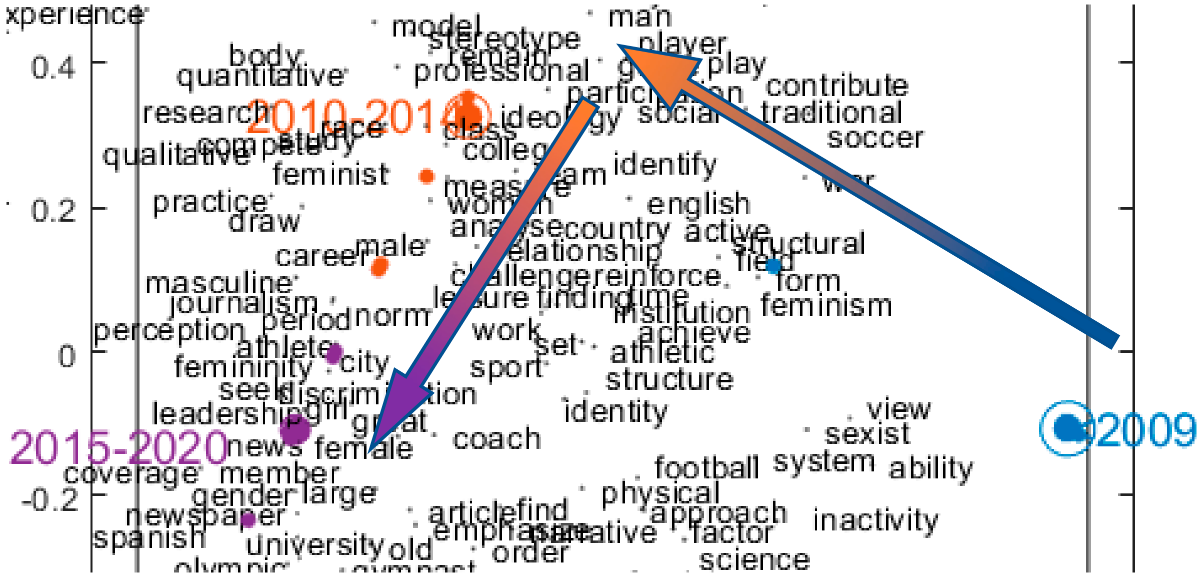

Figure 6. Evolution of the literature of inequalities in women's sport.

Finally, Figure 7 shows the thematic areas that we could identify by highlighting the words that have a greater discriminating power. Thus, MANOVA Biplot brings us a solution that is easier to read than CFA or Reinert's method, and shows the evolution of the publications as well as the specific vocabulary of each period.

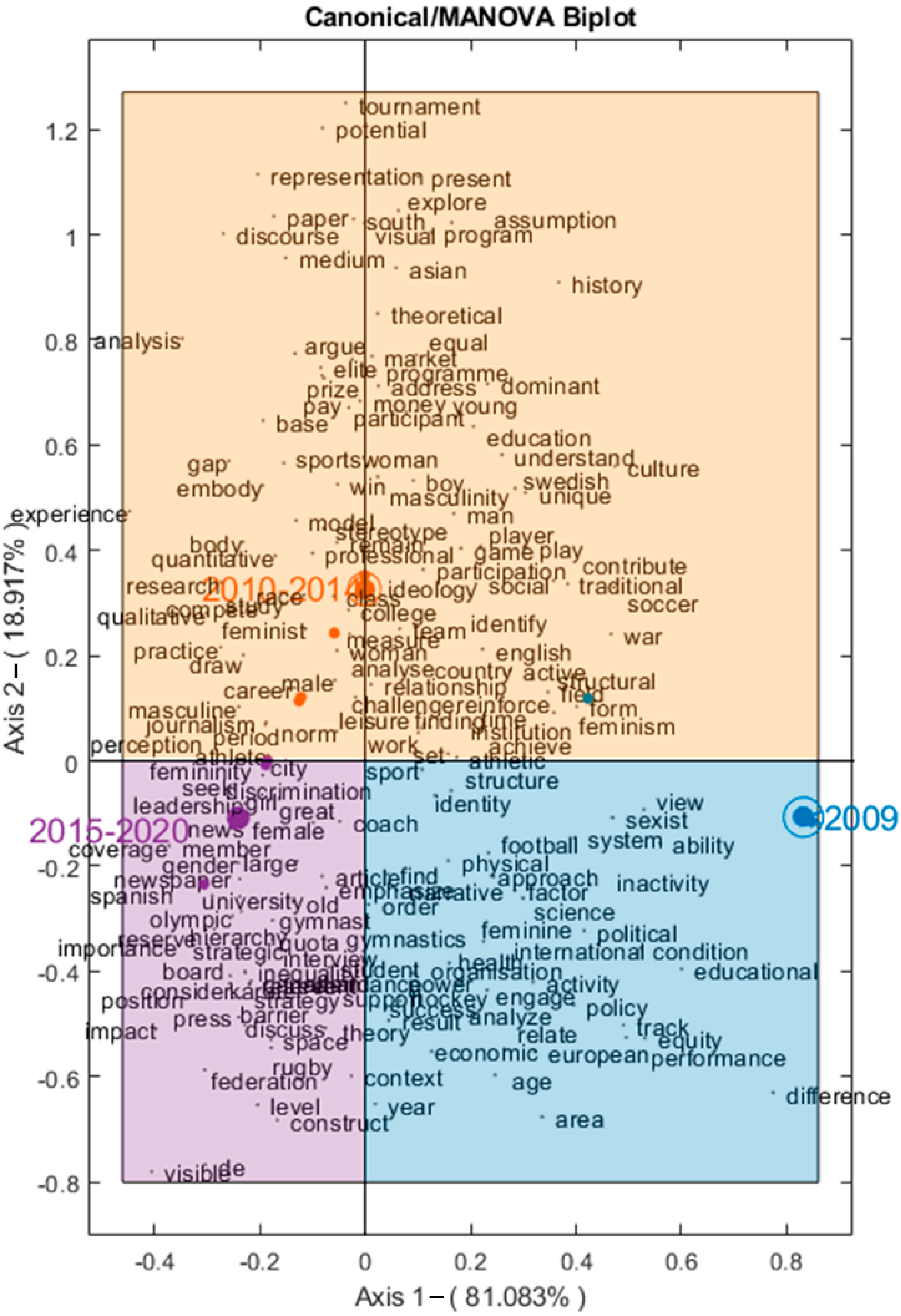

Figure 7. Thematic areas identified by the MANOVA Biplot for the SLR of inequalities in women's sport.

In the same way, if we analyze the results obtained from the MANOVA Biplot analysis applied to the lexical table from the SRL on the professionalization of women's sport, we can identify two major thematic blocks that classify the articles. The two predefined groups served to show a change in the trend in scientific publications. This change took place in 2010. Although the differences are not marked in the analysis of inequalities, there are significant differences between the publications that appeared before 2010 (third and fourth quadrants), and those that appeared onward from that date.

In Figure 8, we can see that these two large groups of words are organized along the second axis. Likewise, we can distinguish two of the keywords that we used in this research very clearly: "woman" in the same direction of axis one and "sport" being located as a variable of the plane (between the two axes). The characterization of the groups therefore lies especially between the positive and negative values of the second axis. Thus, before 2010, in the lower part of the graph, there is vocabulary that mainly uses terms, such as "participation", "participate", "role", "conflict", "critical", "femineity", "balance", "promotional", "defense", "disorder", "mother", "menstrual", "leisure", "management". This reveals a concern of scientists on the participation of women in sports, but not in regards to the professionalization of the practice. It was an era of demands for the inclusion of women in the world of sport, a period of structuring of women's sport.

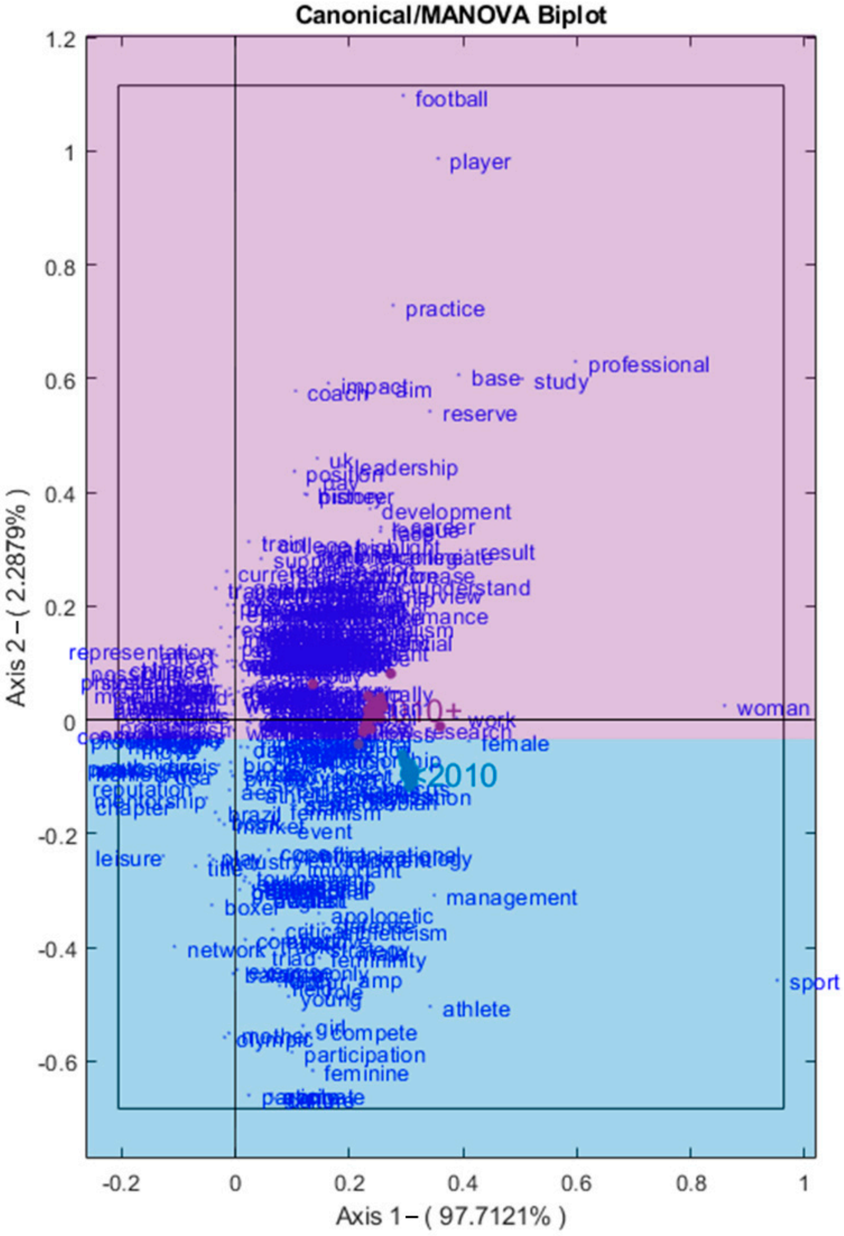

Figure 8. MANOVA Biplot for the SLR of the professionalization of women's sport.

From 2010, the content of articles on women's sport and its professionalization changed towards a discourse focused on the challenges concerning the organization of professional sport for women. We therefore discover words, such as "coach", "study", "professional", "position", "development", "career", "institution", "employment". These words are characteristic of a practice that is developing, but above all which is becoming more and more professional. In this sense, we no longer speak of participation, but of career, specific positions and development. We then see that the themes and issues evolve at the same time as the practice. It is important to underline that certain sports, such as football, have had an important weight in this process of professionalization. "Football", "player" and "practice" are words that appear very clearly in the MANOVA Biplot graphic. Of course, football has been one of the levers for women to make women's sporting practice visible, as well as one of the most important vectors for the professionalization of women in sport.

However, Reinert's method and CFA (Figures 9 and 10) are not able to discriminate between before and after this transition of women's sport appeared at 2010. In contrast, Reinert proposes four categories of clustered data. The first one talks about women's sport professionalization; the second one shows the appearance of content on the economic aspects of sport; and the third one covers the methodology used by authors.

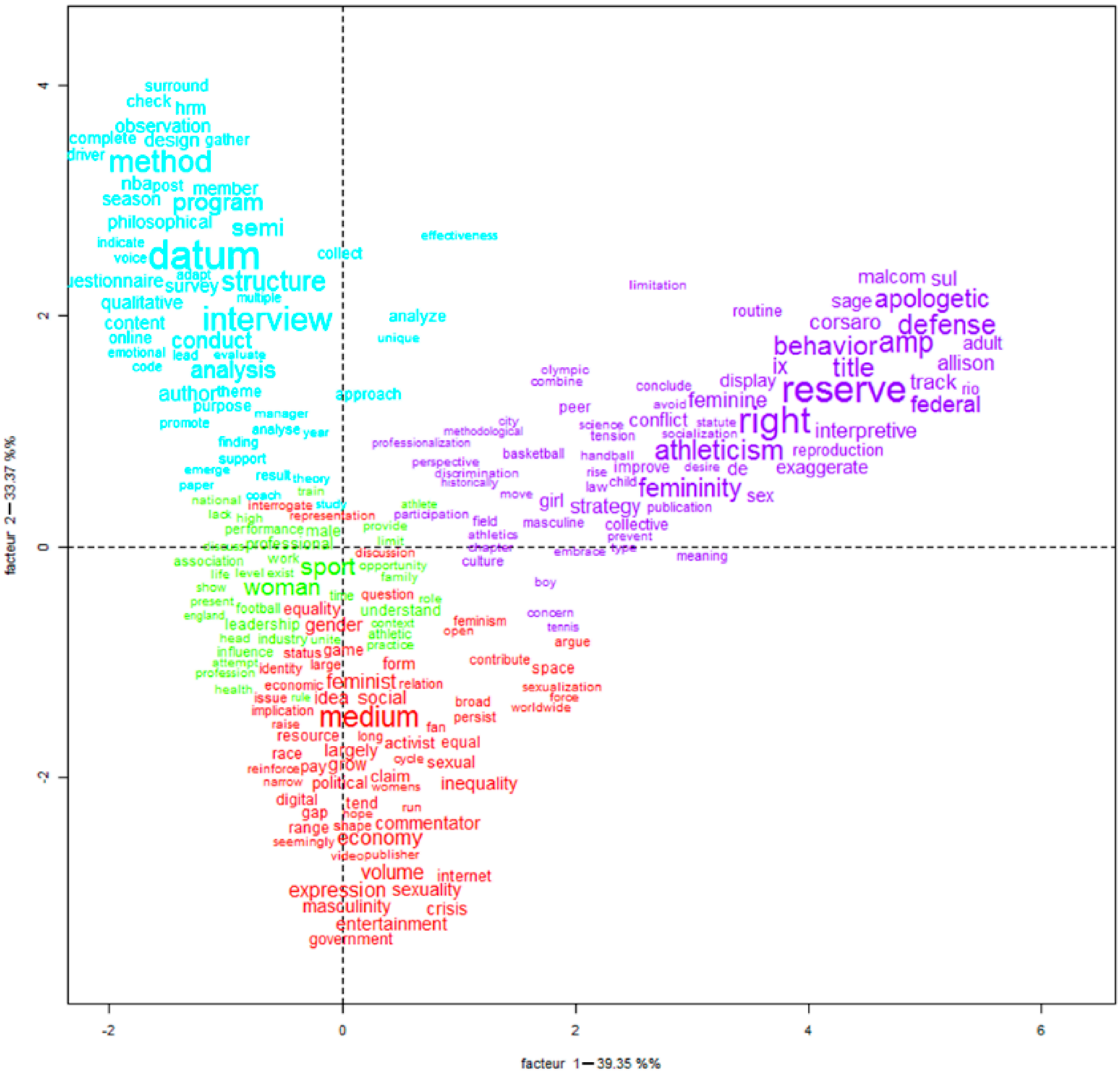

Figure 9. SRL of the professionalization of women's sport using CFA.

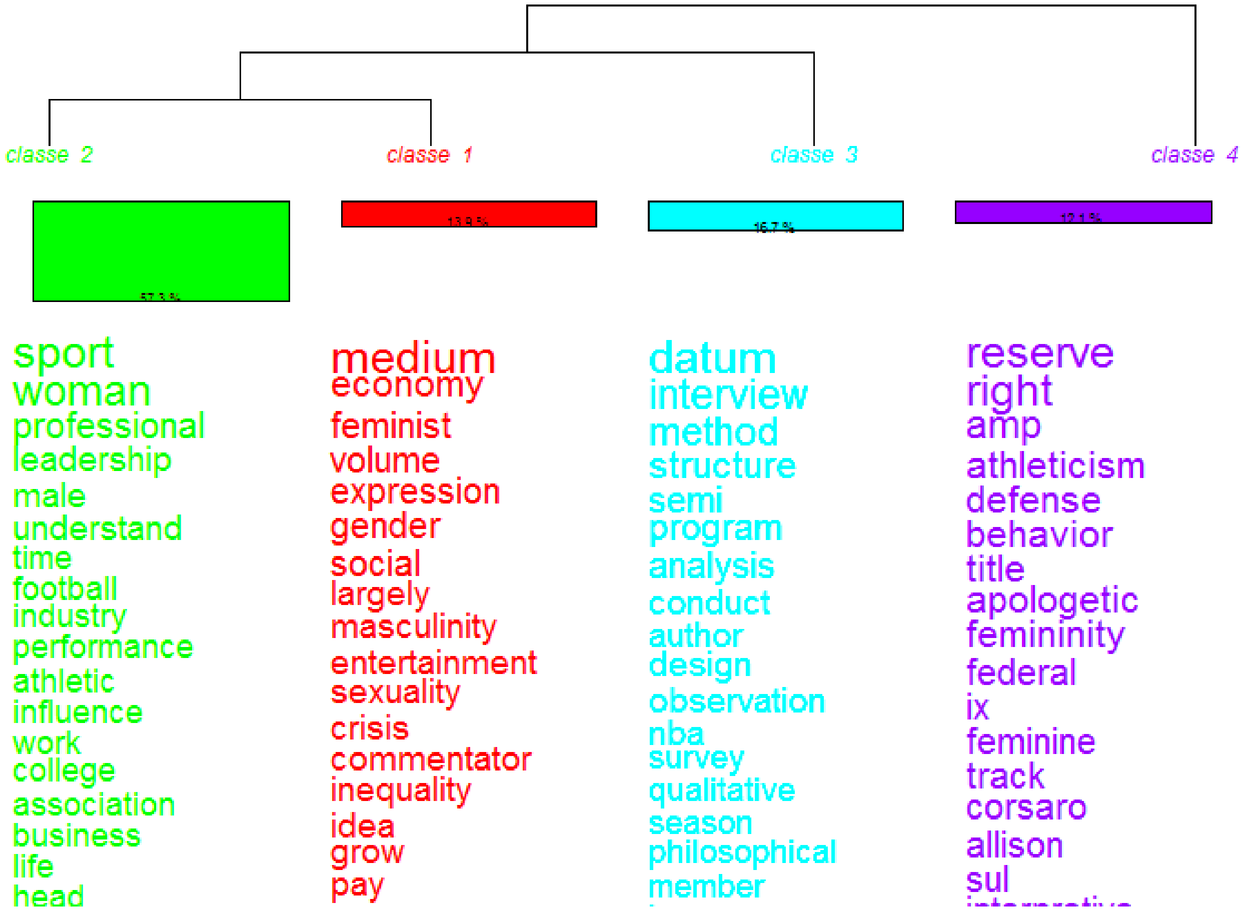

Figure 10. SRL of the professionalization of women's sport using Reinert's method.

6.2. Example 2: SRL in Health Sciences (Low Back Pain Prevention)

The same procedure was used to describe the evolution of the publications concerning low back pain prevention. In this more specific case, we find 141 articles using the keywords "low back pain education prevention", in which 77 articles were included according to the subject.

The CFA (see Figure 11) shows us a graphical solution, displaying the cloud of words in which we can see three different vocabularies. On one hand, we have two groups of words at the positives values of the abscissa axis, distinguishing the differences between those groups at the positive (first group) and negative (second group) values of the ordinate axis. Those two groups refer to the physical (first quadrant) and psychological (fourth quadrant) aspects. On the other hand, the last group is placed throughout the negative values of the abscissa axis. This last group includes the social and educational dimension of the literature of low back pain prevention. As we can see, this CFA solution highlights the three dimensions of health, according to its official definition.

Reinert's method (see Figure 12) describes a more developed solution, with five classes (topics) highlighting the principle words of each one. According to these results, we can expect a classification of publications on low back pain prevention. We will first find a profile concerning the medical aspects of the treatment of low back pain; a second one on the prevention of this illness and the evidence of physical activity efficacy; a third one that concerns the publications on professional activity and the effects of injury at a job; the fourth one concerns physical aspects of low back pain and the relation between age and health; and the last one concerns methodological aspects of a systematic review.

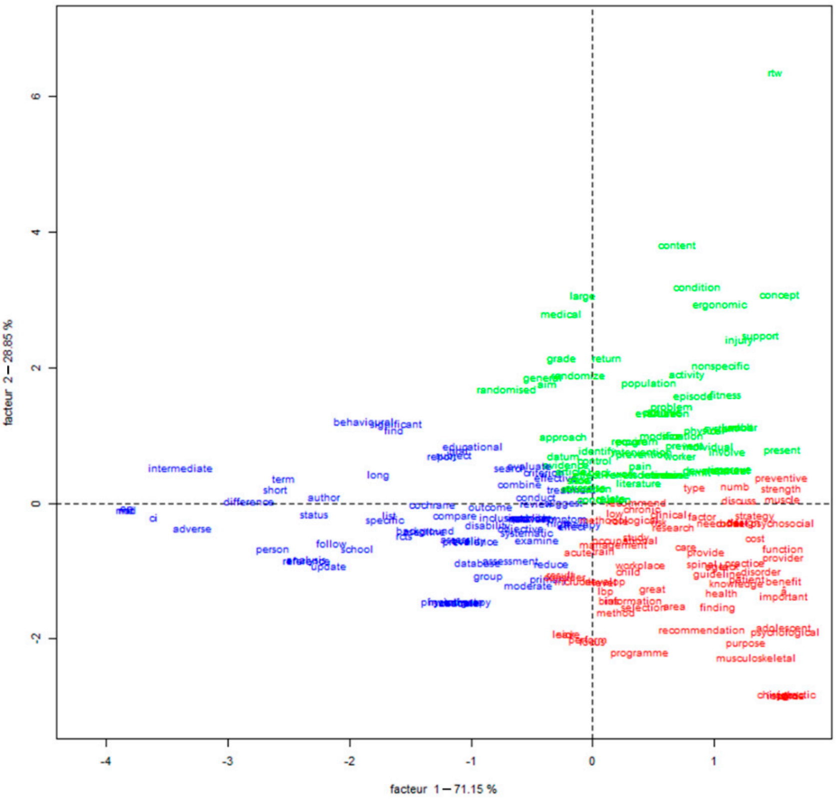

Figure 11. SRL in health sciences (low back pain prevention) using CFA.

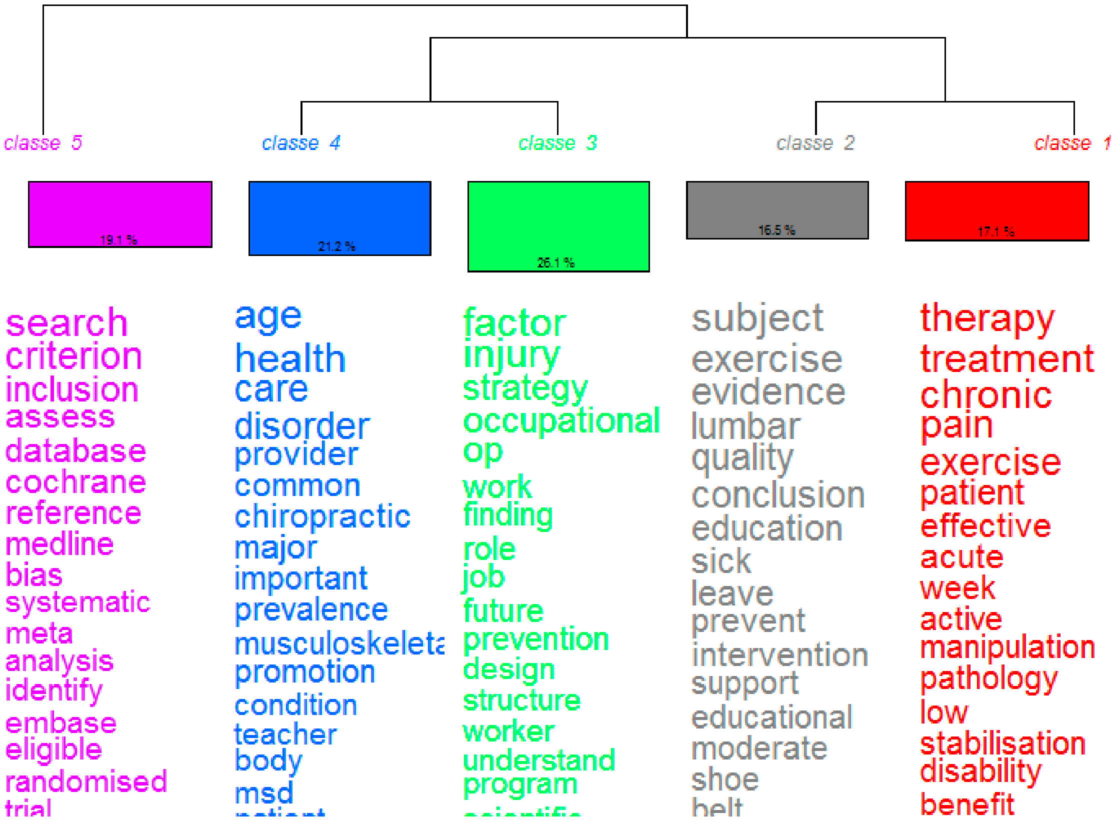

Figure 12. SRL in health sciences (low back pain prevention) using Reinert's method.

The analysis using MANOVA Biplot returns a solution, in which we can see the evolution of publications and how authors have changed their interest on low back pain prevention. As Figure 13 shows, almost every period is well distinguished on the 1–2 plane of the MANOVA Biplot. The oldest publications are placed into the first quadrant with a vocabulary composed by words, such as “design”, “job”, “offer”, “injury”, “workplace”. We can associate those publications with the first class of the Reinert’s method, revealing the authors’ interest about work and prevention of illness at the workplace. However, this interest has evolved into a more methodological type of publication (for example, meta-analysis or systematic review) in 2000–2004. Later, authors have been interested once again on musculoskeletal aspects of low back pain, such as 1995–1999 and previously. Finally, it is in the last two decades that authors have been interested in education and prevention programs. This interpretation allows us to identify that it is particularly in the last years that authors have been worried about educational aspect of prevention of low back pain.

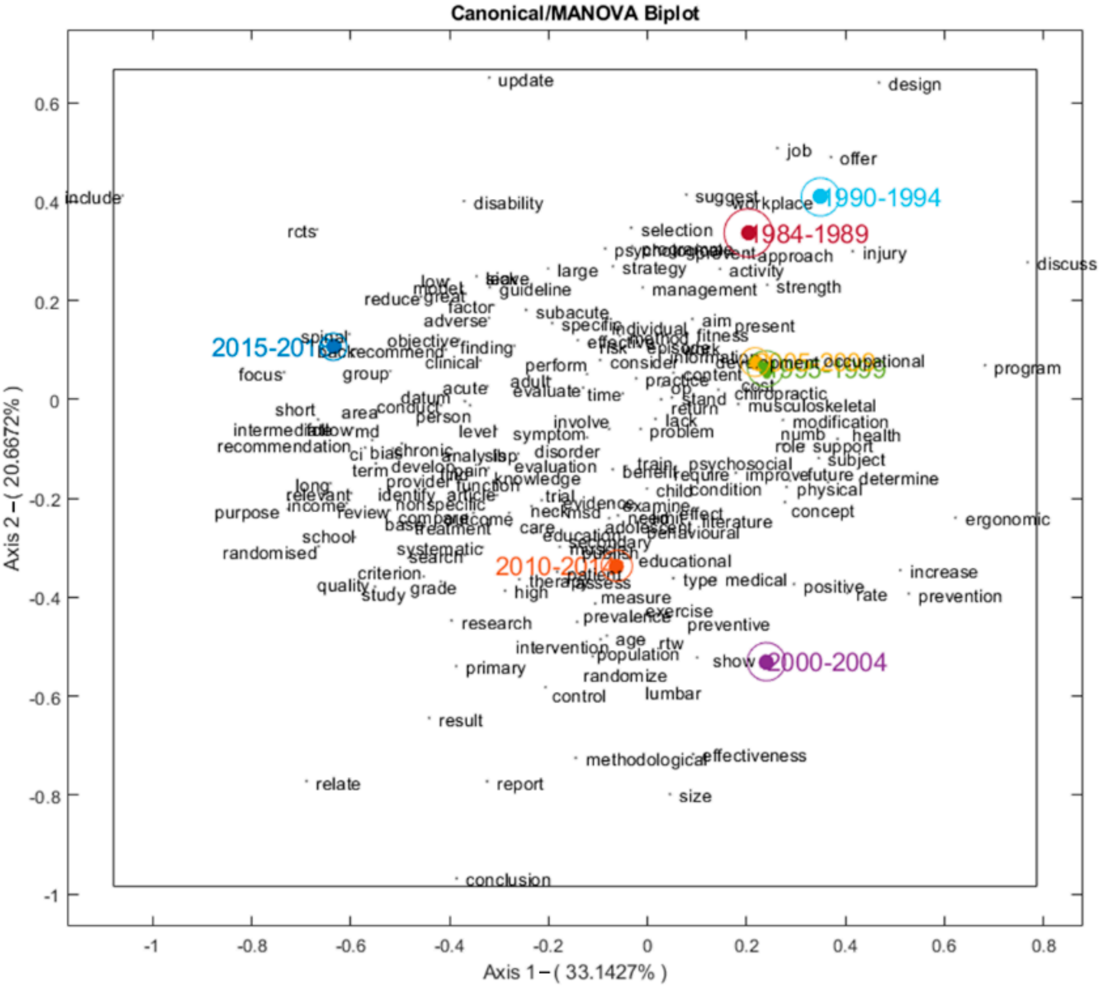

Figure 13. SRL in health sciences (low back pain prevention) using MANOVA Biplot.

These publications could be found on databases, but similarly as we can see in Figure 14, Web Of Science publishes differently from Scopus or Pubmed. In fact, the publications, in which the vocabulary is focused on education are especially published on Web of Science database. However, there are no significant differences between Pubmed and Scopus.

6.3. Example 2: SRL in Multidisciplinary Area (Text Mining)

Finally, we will finish this section by entering a last example, based on a multidisciplinary topic, text mining. The systematic review of literature on text mining for the last five years revealed that there are no differences according to the year of publication (see

Figure 15). This means that authors have been writing about the same topics. However, a new topic appeared in 2020 and continues into 2021: COVID-19. In this way, SRL shows the interest in using text mining for the new pandemic analysis. Others variables, such as area of knowledge or country may be used in order to observe the differences in publications.

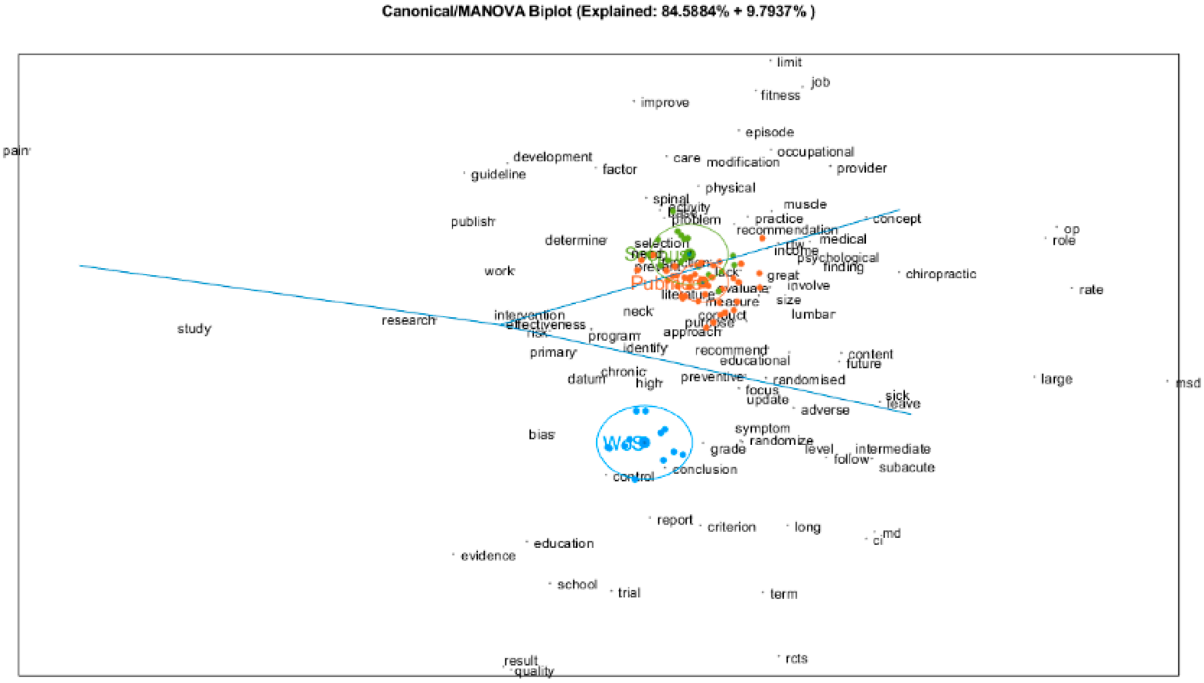

Figure 14. SRL in health sciences (low back pain prevention) using MANOVA Biplot.

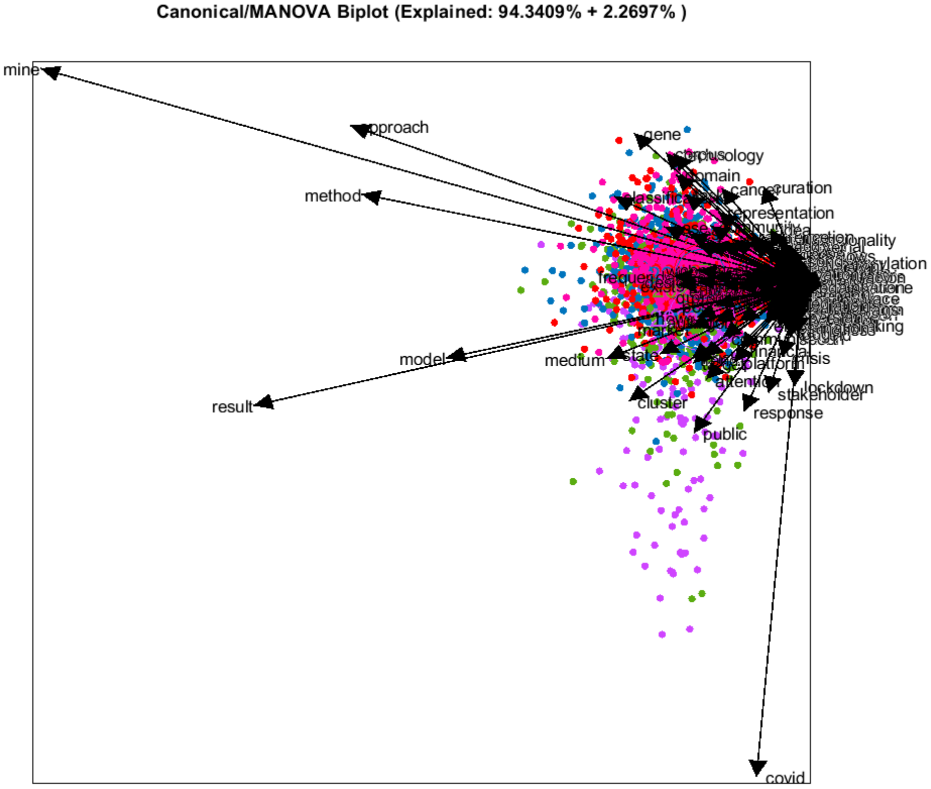

Figure 15. SRL of text mining using MANOVA Biplot.

Finally, the Reinert method reveals five topics in the publications (see Figure 16). The first one regroups the vocabulary of the online commerce, where text mining is used to analyze customer opinions. This first class is associated with the third and fourth ones, that concern articles talking about literature analysis and policy development, respectively. The second and the last one regroups natural language recognition for biomedical (class 2) issues and technical aspects of text mining (class 5).

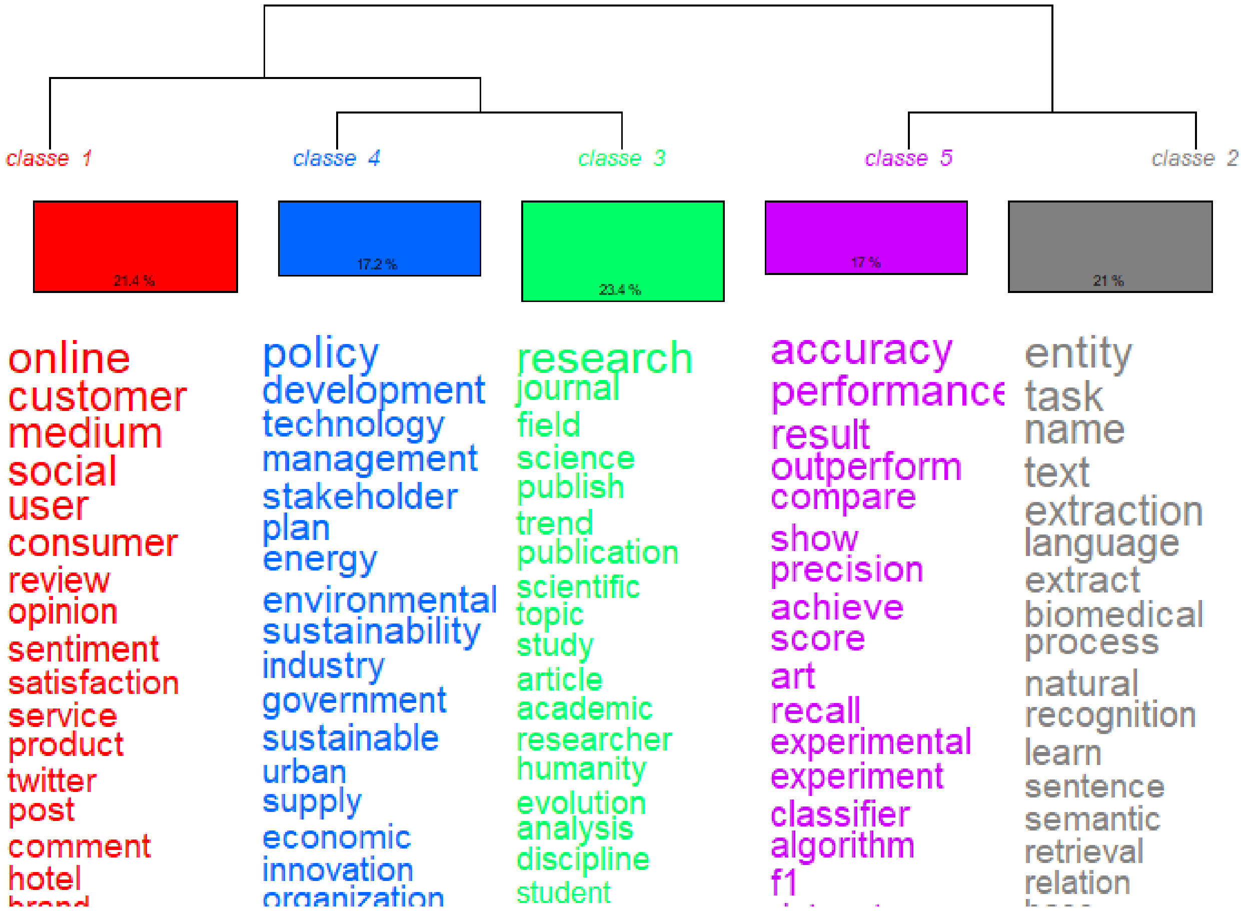

Figure 16. SRL of text mining using Reinert's method.

Ultimately, we can see that from the MANOVA Biplot, if we follow the steps described below, it is possible to quickly identify the main thematic lines in the literature on each topic. In addition, we also note that thanks to the vocabulary shared by articles from the same period, bibliographic research can be greatly simplified. The articles, although each retain their autonomy and originality, are grouped together in their projections on the lesser dimensional plane of the biplot, showing us the (di) similarities between them. This allows us to select representative articles from each group, in order to perform additional verification, allowing us to confirm our interpretations.

This analysis, even if it may be complemented with Reinert's method, returns a complete solution that made reasonably easy in its interpretation (topics, vocabulary, evolution, etc.). Let us not forget that this is a method that is still in the process of development. It may be improved by analyzing the set of articles and not only the abstracts. However, a simple reading of the articles that are closest to the centroids of each group made it possible to validate the reading of the MANOVA Biplots, by making the technique useful not only to identify the main thematic lines but also to summarize the content of a large number of articles in analyzing only a few key articles.

The lexical table construction using the characterization value and the biplot methods improves the inclusion of epistemological considerations in the text mining. The specific vocabularies and the interpretation of topics from the more characteristic words facilitates the interpretation of the meaning of articles in a graphical solution.

7. Conclusions

Text mining technology has considerably enriched traditional knowledge processing tools, in particular, in improving their predictive potential. As presented, this complex tool has undergone numerous and progressive developments. It mainly highlights its usefulness in describing and identifying the most salient textual data [57].

On the proposed technique, we can conclude that it facilitates the systematic review of literature through the use of text mining. In this sense, the MANOVA Biplot, thanks to its discriminating power, has been very useful in simplifying the SLR by categorizing articles that have appeared in different time periods. The interpretation of the results is consistent with reading a small number of articles representative of the periods. The nature of the analyzes, as well as the properties of the lexical table construction process make this method highly customizable. Scientists will be able to target not only the keywords most relevant to their research, but also the bibliographic variables they wish to highlight for an in-depth analysis of the publication.

Today, there are other methods of quantitative analysis of textual or bibliographic content that should be compared, in terms of the qualitative results obtained with respect to the technique presented [58]. Regarding this technique, but also the others, it would be interesting to test the results between different users using the same processing and decryption algorithm, in order to understand the differences in interpretation and the qualitative assessment obtained.

The ability of text mining to generate knowledge from mathematical algorithms is clearly shown in this article. Text mining that is proposed here goes a little further in incorporating the meaning of the texts that are analyzed, one of the challenges that has been highlighted for the future of this type of analysis. In this sense, the method has a fertile ground for the development of protocols, capable of improving epistemological compatibility between quantitative and qualitative research processes, such as the incorporation of qualitative codes [26], improvement of matrix correction factors, among others.

Although the results are satisfactory, the literature analysis was carried out only from the abstracts of articles published on the databases. Although these represent the substance and the essential conclusions of the articles, a more in-depth analysis using all of the text contained in the articles could be considered to refine the interpretations. On this aspect, comparisons would be interesting between the conclusions obtained from the summaries of the articles and the contents of the articles without summaries. However, this present work is much more complex, given the difficulty of obtaining the documents, linguistic differences and the computer processing of the texts. Improved search engines and systems will undoubtedly make it easier and more efficient for scientists to access. Indeed, the comparison of the content of bibliographic databases, according to the indexed journals and certain keywords tested, sometimes express distinct tendencies of interpretation. The orientation of bibliographic search metasearch engines could prove to be very useful in saving time and probably in the quality of analysis.

Author Contributions: Conceptualization, D.C.-J.; methodology, D.C.-J.; software, D.C.-J.; validation, D.C.-J.; formal analysis, D.C.-J.; investigation, D.C.-J.; resources, D.C.-J. and P.C.; data curation, D.C.-J.; writing—original draft preparation, D.C.-J.; writing—review and editing, D.C.-J. and P.C.; visualization, D.C.-J.; supervision, D.C.-J.; project administration, D.C.-J. and P.C.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

References

1. Kar, A.K.; Dwivedi, Y.K. Theory building with big data-driven research—Moving away from the "What" towards the "Why". *Int. J. Inf. Manag.* **2020**, *54*, 102205. [CrossRef]
2. Antons, D.; Grünwald, E.; Cichy, P.; Salge, T. The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. *R&D Manag.* **2020**, *50*, 329–351. [CrossRef]
3. Törnberg, P.; Törnberg, A. The limits of computation: A philosophical critique of contemporary Big Data research. *Big Data Soc.* **2018**, *5*. [CrossRef]
4. Greenacre, M. Contribution biplots. *J. Comput. Graph. Stat.* **2013**, *22*, 107–122. [CrossRef]
5. Gabriel, K.R. MANOVA biplots for twoway contingency tables. In *Recent Advances in Descriptive Multivariate Analysis*; Krzanowski, W., Ed.; Clarendon Press: Oxford, UK, 1995; pp. 227–268.
6. Gower, J.; Lubbe, S.; le Roux, N. *Understanding Biplots*; Wiley: Chichester, UK, 2011.
7. Luhn, H.P. The automatic creation of literature abstracts. *IBM J. Res. Dev.* **1958**, *2*, 159–165. [CrossRef]
8. McCarthy, J. Epistemological problems of artificial intelligence. In *Readings in Artificial Intelligence*; Elsevier: Amsterdam, The Netherlands, 1981; pp. 459–465. [CrossRef]
9. Ng, C.; Alarcon, J. Text mining. In *Artificial Intelligence in Accounting: Practical Applications*; Routledge: Oxfordshire, UK, 2020; pp. 46–70.
10. Avasthi, S.; Chauhan, R.; Acharjya, D.P. Techniques, applications, and issues in mining large-scale text databases. In *Advances in Information Communication Technology and Computing*; Springer: New York, NY, USA, 2020; pp. 385–396.
11. Chambua, J.; Niu, Z. Review text based rating prediction approaches: Preference knowledge learning, representation and utilization. *Artif. Intell. Rev.* **2021**, *54*, 1171–1200. [CrossRef]
12. Malhotra, N.K.; Rush Charles, B.; Uslay, C. Correspondence analysis. Methodological perspectives, issues, and applications. In *Review of Marketing Research (Review of Marketing Research, Vol. 1)*; Emerald Group Publishing Limited: Bingley, UK, 2005; pp. 285–316.
13. Benzécri, J.P. *L'Analyse des Données: L'Analyse des Correspondances*; Dunod: Paris, Fance, 1973.
14. Benzécri, J.P. Statistical analysis as a tool to make patterns emerge from data. In *Methodologies of Pattern Recognition*; Watanabe, S., Ed.; Academic Press: New York, NY, USA, 1969; pp. 35–74.
15. Benzécri, J.P. *Correspondence Analysis Handbook*; Dekker: New York, NY, USA, 1992.
16. Lebart, L.; Morineau, A.; Warwick, K. *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*; Wiley: New York, NY, USA, 1984.
17. Lebart, L.; Salem, A. *Statistique Textuelle*; Dunod: Paris, France, 1994.
18. Lebart, L.; Salem, A.; Berry, L. *Exploring Textual Data*; Springer: Dordrecht, The Netherlands, 2011.
19. Reinert, M. Classification descendante hiérarchique et analyse lexicale par contexte: Application au corpus des poésies d'A. Rimbaud. *Bull. Méthodol. Sociol.* **1987**, *13*, 53–90. [CrossRef]
20. Reinert, M. Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval. *Bull. Méthodol. Sociol.* **1990**, *26*, 24–54. [CrossRef]
21. Reinert, M. Proposition d'une méthodologie d'analyse des données séquentielles. *Bull. la Société Française pour l'Etude du Comport. Anim.* **1991**, *1*, 53–60.
22. Reinert, M. Système Alceste: Une méthodologie d'analyse des données textuelles. In *JADT 1990*; Polytechnic University of Catalonia: Barcelona, Spain, 1992; pp. 144–161.
23. Osuna, Z. Contribuciones al Análisis de Datos Textuales. Ph.D. Thesis, Universidad de Salamanca, Salamanca, Spain, 2006.
24. Gabriel, K.R. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **1971**, *58*, 453–467. [CrossRef]
25. Dalud-Vincent, M. Alceste comme outil de traitement d'entretiens semi-directifs: Essai et critiques pour un usage en sociologie. *Lang. Société* **2011**, *135*, 9. [CrossRef]
26. Caballero-Julia, D.; Vicente, M.P.; Galindo, M.P. Grupos de discusión y HJ-biplot: Una nueva forma de análisis textual. *RISTI Rev. Ibérica Sist. Technol. Inf.* **2014**, *2*, 19–36. [CrossRef]
27. Galindo, M.P.; Cuadras, C.M. *Una Extensión del Método Biplot y su Relación con Otras Técnicas*; Universidad de Barcelona: Barcelona, Spain, 1986.
28. Galindo, M.P. Contribuciones a la Representación Simultánea de Datos Muldimensionales. Ph.D. Thesis, Universidad de Salamanca, Salamanca, Spain, 1985.
29. Galindo, M.P. An alternative for simultaneous representation: HJ-biplot. *Questiió* **1986**, *10*, 13–23.
30. Martin, A.; Adelé, S.; Reutenauer, C. Stratégies du voyageur: Analyse croisée d'entretiens semi-directifs. In Proceedings of the 13ème Journées internationales d'Analyse statistique des Données Textuelles, Nice, France, 13–15 June 2016.
31. Heiden, S.; Magué, J.-P.; Pincemin, B. TXM: Une plateforme logicielle open-source pour la textométrie—Conception et développement. In Proceedings of the 10th International Conference on the Statistical Analysis of Textual Data—JADT 2010, Rome, Italy, 6–11 June 2010.
32. Ratinaud, P. Amélioration de la précision et de la vitesse de l'algorithme de classification de la méthode Reinert dans IRaMuTeQ. In Proceedings of the 14th International Conference on Statistical Analysis of Textual Data, Rome, Italy, 12–14 June 2018.
33. Bécue-Bertaut, M. *Analyse Textuelle Avec R*; Presses Universitaires de Rennes: Rennes, France, 2018.
34. Paveau, M.-A. L'alternative quantitatif/qualitatif à l'épreuve des univers discursifs numériques. *Corela* **2014**. [CrossRef]
35. Merriam, S.; Tisdell, E. *Qualitative Research: A Guide to Design and Implementation*; Jossey-Bass: San Francisco, CA, USA, 2015.

36. Dumez, H. *Comprehensive Research: A Methodological and Epistemological Introduction to Qualitative Research*; Business School Press: Copenhagen, Denmark, 2016.
37. Wu, X.; Zhu, X.; Wu, G.Q.; Ding, W. Data mining with big data. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 97–107. [CrossRef]
38. Kune, R.; Konugurthi, P.K.; Agarwal, A.; Chillarige, R.R.; Buyya, R. The anatomy of big data computing. *Softw. Pract. Exp.* **2016**, *46*, 79–105. [CrossRef]
39. Snyder, H. Literature review as a research methodology: An overview and guidelines. *J. Bus. Res.* **2019**, *104*, 333–339. [CrossRef]
40. The World Bank. Scientific and Technical Journal Articles. 2021. Available online: https://data.worldbank.org/indicator/IP.JRN.ARTC.SC?end=2018&start=2000&view=chart (accessed on 2 June 2021).
41. Boyd, D.; Crawford, K. Critical questions for big data. *Inf. Commun. Soc.* **2012**, *15*, 662–679. [CrossRef]
42. Favaretto, M.; De Clercq, E.; Schneble, C.O.; Elger, B.S. What is your definition of big data? Researchers' understanding of the phenomenon of the decade. *PLoS ONE* **2020**, *15*, e0228987. [CrossRef]
43. Munn, Z.; Peters, M.D.J.; Stern, C.; Tufanaru, C.; McArthur, A.; Aromataris, E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med. Res. Methodol.* **2018**, *18*, 1–7. [CrossRef]
44. Higgins, J.; Green, S. (Eds.) *Cochrane Handbook for Systematic Reviews of Interventions*; 5.1.0; The Cochrane Collaboration: London, UK, 2011.
45. Pranckutė, R. Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. *Publications* **2021**, *9*, 12. [CrossRef]
46. Clarivate. Web of Science Platform. Available online: https://clarivate.libguides.com/webofscienceplatform/coverage (accessed on 9 June 2021).
47. Elsevier. Content Coverage Guide. 2020. Available online: https://www.elsevier.com/__data/assets/pdf_file/0007/69451/Scopus_ContentCoverage_Guide_WEB.pdf (accessed on 9 June 2021).
48. Elsevier. Scopus. Available online: https://www.scopus.com/ (accessed on 9 June 2021).
49. Caballero-Julia, D. *El HJ-Biplot como Herramienta en el Análisis de Grupos de Discusión*; Repositorio Institucional Gredos de la Universidad de Salamanca: Salamanca, Spain, 2011.
50. Osuna, Z.; Galindo-Villardon, M.P.; Martin-Vallejo, J. Análisis estadístico de datos textuales. Aplicación al estudio de las declaraciones del Libertador Simón Bolívar. *Aled Rev. Latinoam. Estud. del Discurso* **2004**, *4*, 55–62. [CrossRef]
51. Gabriel, K.R. Analysis of meteorological data by means of canonical decomposition and Biplots. *J. Appl. Meteorol.* **1972**, *11*, 1071–1077. [CrossRef]
52. Amaro, I.R.; Vicente, J.L.; Galindo, M.P. MANOVA biplot para arreglos de tratamientos con dos factores basado en modelos lineales generales multivariantes. *Interciencia* **2004**, *29*, 26–32.
53. Caballero-Julia, D.; Galindo, M.P.; Garcia, M.-C. JK-meta-biplot y STATIS dual como herramientas de análisis de tablas textuales múltiples. *RISTI—Rev. Iber. Sist. e Tecnol. Inf.* **2017**, *25*, 18–33. [CrossRef]
54. Vicente, J.L.; Galindo, M.P.; Avila, C.; Fernandez, M.J.; Martín, J.; Bacala, N. JK-META-BIPLOT: Una alternativa al método statis para el estudio espacio temporal de ecosistemas. In Proceedings of the Conferencia Internacional de Estadística en Estudios Medioambientales, Universidad de Cádiz, Cádiz, Spain, 21–23 November 2001.
55. Varas, M.J.; Vicente, S.; Molina, E.; Vicente, J.L. Role of canonical biplot method in the study of building stones: An example from spanish monumental heritage. *Environmetrics* **2005**, *16*, 405–419. [CrossRef]
56. Vicente, J.L. *MULTBIPLOT: A Package for Multivariate Analysis Using Biplots*; Departamento de Estadística, Universidad de Salamanca: Salamanca, Spain, 2014.
57. Gallagher, R.J.; Frank, M.R.; Mitchell, L.; Schwartz, A.J.; Reagan, A.J.; Danforth, C.M.; Dodds, P.S. Generalized word shift graphs: A method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Sci.* **2021**, *10*, 4. [CrossRef]
58. Nelson, L.K.; Burk, D.; Knudsen, M.; McCall, L. The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociol. Methods Res.* **2021**, *50*, 202–237. [CrossRef]