mathematics

MDPI

Article

HJ-Biplot as a Tool to Give an Extra Analytical Boost for the Latent Dirichlet Assignment (LDA) Model: With an Application to Digital News Analysis about COVID-19

Luis Pilacuan-Bonete [1,2,*], Purificación Galindo-Villardón [1,3,4] and Francisco Delgado-Álvarez [1]

[1] Department of Statistics, University of Salamanca, 37008 Salamanca, Spain; pgalindo@usal.es (P.G.-V.); j.delgado@usal.es (F.D.-Á.)
[2] Faculty of Industrial Engineering, Universidad de Guayaquil, Guayaquil 090514, Ecuador
[3] Escuela Superior Politécnica del Litoral, Escuela Superior Politécnica del Litoral (ESPOL), Centro de Estudios e Investigaciones Estadísticas, Campus Gustavo Galindo, Km. 30.5 Via Perimetral, Guayaquil P.O. Box 09-01-5863, Ecuador
[4] Centro de Gestión de Estudios Estadísticos, Universidad Estatal de Milagro (UNEMI), Ciudadela Universitaria Km. 1.5 vía al Km 26, Guayas 091050, Ecuador
* Correspondence: luis.pilacuanb@usal.es; Tel.: +593-981105994

Abstract: This work objective is to generate an HJ-biplot representation for the content analysis obtained by latent Dirichlet assignment (LDA) of the headlines of three Spanish newspapers in their web versions referring to the topic of the pandemic caused by the SARS-CoV-2 virus (COVID-19) with more than 500 million affected and almost six million deaths to date. The HJ-biplot is used to give an extra analytical boost to the model, it is an easy-to-interpret multivariate technique which does not require in-depth knowledge of statistics, allows capturing the relationship between the topics about the COVID-19 news and the three digital newspapers, and it compares them with LDAvis and heatmap representations, the HJ-biplot provides a better representation and visualization, allowing us to analyze the relationship between each newspaper analyzed (column markers represented by vectors) and the 14 topics obtained from the LDA model (row markers represented by points) represented in the plane with the greatest informative capacity. It is concluded that the newspapers El Mundo and 20 M present greater homogeneity between the topics published during the pandemic, while El País presents topics that are less related to the other two newspapers, highlighting topics such as t_12 (Government_Madrid) and t_13 (Government_millions).

Keywords: SARS-CoV-2; COVID-19; HJ-biplot; latent Dirichlet assignment; LDA

MSC: 62H35

Citation: Pilacuan-Bonete, L.; Galindo-Villardón, P.; Delgado-Álvarez, F. HJ-Biplot as a Tool to Give an Extra Analytical Boost for the Latent Dirichlet Assignment (LDA) Model: With an Application to Digital News Analysis about COVID-19. *Mathematics* **2022**, *10*, 2529. https://doi.org/10.3390/math10142529

Academic Editor: Andrea De Gaetano

Received: 1 June 2022
Accepted: 24 June 2022
Published: 20 July 2022

1. Introduction

Humanity is suffering from a pandemic caused by the SARS-CoV-2 virus with more than 500 million people affected and almost six million deaths to date. This tragic situation is causing opinions and information related to SARS-CoV-2 (COVID-19) to be written in all available media: print and digital press, social networks, web pages, forums, etc. A technological resource available for the analysis of this web information is the textual analysis of content, which is being used regularly in all types of environments, including web environments. Much research has been carried out on this type of analysis, for different applications, such as studying publications on social networks [1], analyzing the marketing management of companies [2], establishing worker profiles from LinkedIn [3], analysis of scientific literature [4], and analyzing effects on health workers through Twitter posts [5], among others.

As of 27 May 2022, there were 528,431,653 confirmed cases of the SARS-CoV-2 virus registered in the world according to the World Health Organization [6]. This disease was

first identified in hospitalized patients in Wuhan, China, in December 2019 [7] and is associated with symptoms of severe pneumonia, causing fever, cough, and respiratory failure [8], being able to cause the death of the infected patient. The COVID-19 disease was declared a pandemic by the WHO on 11 March 2020, and is affecting not only human health but also the world economy [9], generating high interest in people searching for information in credible media such as newspapers. In Spain, the first case of infection by the virus was registered on the island of La Gomera on 31 January 2020 [10] and as of 27 May of the same year, 12,326,264 people had been infected [6].

According to the Association for Media Research (AIMC), the three newspapers in Spain with the highest number of daily readers are El País, El Mundo, and 20 Minutos [11] with more than 570k readers per day each. The news published by the newspapers can influence the public opinion of the readers [12]. Multiple studies have analyzed this influence on the readers of the news published in various newspapers, such as in politics [13], consumption [14], and discrimination against criminals [15]. The world is currently experiencing dark times due to the pandemic caused by SARS-CoV-2, also known as the 'coronavirus', making the study of the publications generated on this subject especially relevant.

For the textual analysis of the information available in cyberspace, different techniques are applied to consist of obtaining information from the web, known as 'web scraping', which allows the extraction of part or all the data from web pages, written in different formats such as XML and HTML, among others [16]. These extracted data can be subjected to different transformation processes of semantic and syntactic information, through natural language processing (NLP) techniques for the formulation of text corpus [17].

The information thus extracted is subjected to different textual analysis techniques to analyze this text corpus, such as term frequency analysis (TF), which allows statistical interpretation of the specificity of the term and its application in retrieval [18]. A quantitative corpus is generated that describes the words as variables and the n documents extracted as individuals, to represent the textual fragments as a linear equation. These corpuses are structured following different semantic analysis techniques. Among these techniques are those of latent semantic analysis (LSA), which uses singular value decomposition (SVD) for the segmentation of the corpus matrix, applying the statistical basis of the co-occurrence of words in the corpus and ignoring the grammatical structure [19]. Another methodology used is the latent Dirichlet allocation (LDA), consisting of a three-level hierarchical Bayesian probabilistic model, which allows a collection of observations to be explained as a whole and which shows the similarity between the extracted data [20], another technique is the use of machine learning algorithms, applied in textual studies with MTL multitasking learning models [21], or text sentiment analysis applications based on the synthetic minority over-sampling technique (SMOTE) [22], among others applications.

The text corpus, methods such as the LDAvis [23], a very extended alternative for the representation of the topics, which allows display on a web page using an interactive graphic (scalable vector graphics, SVG). Representation employing a heatmap type has also been proposed [24] allows us to visualize a rearrangement by some set of values, usually the mean of the rows or columns. Another proposed technique is the HJ-biplot method formulated by Galindo [25] in which the rows and columns of a data matrix are represented in the same system of factorial axes; examples can be cited from the works developed to represent bibliometric studies [26], to represent the quality of life discussion groups [27], or to classify the investiture speeches of Spanish rulers [28].

Based on everything mentioned above, the present study aims to generate an HJ-biplot representation of the distribution matrix of the topics on the documents, resulting from an LDA analysis of the news generated regarding the SARS-CoV-2 pandemic in the three generalist Spanish newspapers with the highest number of readers, thus allowing visualization of the relationship for each topic of the news of the newspapers concerning COVID-19.

2. Materials and Methods

For the content analysis of the three most read newspapers in Spain, text mining tools and techniques are used which allow for the generation—through statistical methods—a visualization of the data [29]. Parallelization techniques have been applied in data processing to obtain higher computational performance [30]. A standard text mining process starts from the integration of raw information, coming from different data sources, which is cleaned to eliminate inconsistencies and duplicates that generate noise in the analysis [31]. With the data transformed into a homogeneous format, and through text mining filtering and aggregation techniques, analyses can be carried out where the most interesting existing patterns are identified.

Next, the various techniques applied in each of the processes followed in the analysis will be presented, detailed in Figure 1. These techniques were applied in pre-existing modules for the different analyses in an open-source software R version 3.6.3 [32].

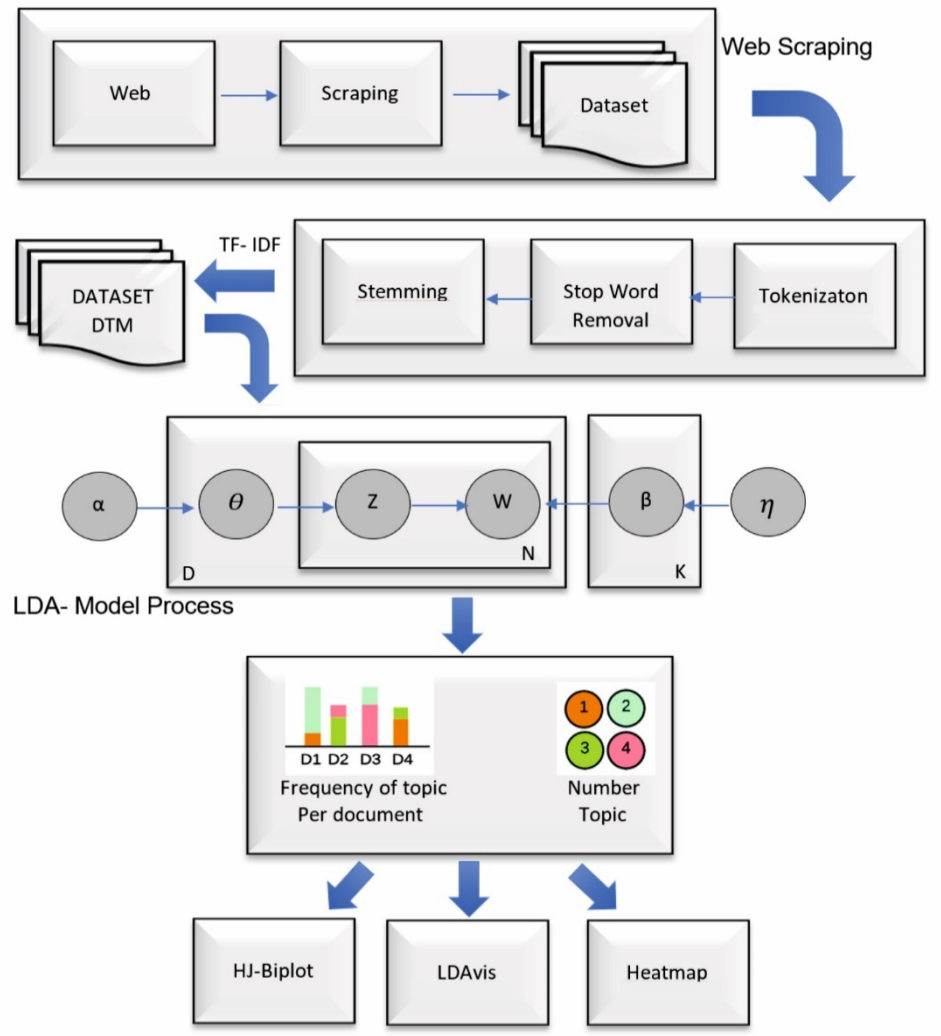

Figure 1. Applied methodology architecture.

2.1. Web Data Extraction

Data extraction techniques from the web have had high growth in recent times. Due to the massification of information on the World Wide Web (WWW), this has become an important global database. Web scraping, used for web content mining, obtains information from the content of web pages, with two basic objectives: extract information to improve search engines and information retrieval fields [33] and analyze and explore information to

gain useful content knowledge [34]. In this content mining technique, opinions, feelings, and emotions are extracted from the text to understand the context of web content [35].

These text analysis processes are widely used in a wide range of applications: there are several studies where these techniques have been applied in the extraction of information from Facebook [36], Twitter [37], web pages [38] in general, and bibliographic sites to obtain academic information, such as expert recommendations [39], among other sites with different study purposes [40]. Rekik [41] concluded that it is important to evaluate the quality of information on websites; for this, firstly, the criteria to carry out the evaluation are defined and regrouped semantically; subsequently, useful information is extracted from them to create a set of data criteria with which to obtain more specific information.

Ferrara [41] detailed how complex it can be to extract information from the web, especially when it comes to unstructured data, or depending on the languages of each page and even the browsers used; but he also described the multiple applications that these techniques have in different fields, taking into account the computational cost of these methodologies.

For the present study, as already indicated, the R software environment will be used, which provides multiple supports for web mining [42], and has different packages and functions that handle data extraction tasks, such as Rcrawler [43]. However, the user must manually manage the content that they want to extract from the URLs—that is, data cannot be obtained automatically—so other types of tools are also required if the extraction process needs to be automated.

In addition, the R package called Rvest (version 0.3.5) has been used, which allows information to be extracted from URLs that are in HTML or XML format [44]. This package, depending on the tasks to be carried out, can be combined with others to incorporate other functionalities. Specifically, in the present study, it was combined with the R packages dplyr (version 0.8.5) and the Base package (version 3.5.0), to extract the unstructured data and subject it to a cleaning and transformation process in character-type text format, to later be submitted to the final content analysis.

2.2. Term Frequency

The methodology called term frequency in the document, or TF (term frequency), found its way into almost all terminology weighting schemes. According to Jones's postulate [18], terms can be said to be words or possibly phrases or word-words; assuming that there are $\boldsymbol{N}$ documents in a collection and that the term $\boldsymbol{t_i}$ appears in $\boldsymbol{n_i}$ of them, then the proposed measure, defined as a weight, will be applied to the term $\boldsymbol{t_i}$, and is described in Equation (1), also known as the inverse document frequency (or IDF), this formulation being one of the most used (Robertson, 2004).

$$\mathrm{idf}(\boldsymbol{t}_i) = log\frac{N}{\boldsymbol{n}_i} \tag{1}$$

The assignment of unique weighted terms properly produces retrieval results superior to those that can be obtained with other text techniques used, depending on the term weighting system chosen [45,46]. In this study, the weighting of terms called TF-IDF is considered, which is a metric where the TF provides a direct estimate of the probability of occurrence of a term, normalized by the total frequency of the document [47] and in which this indicator is multiplied for the IDF, which in turn can be interpreted as the amount of information, given as the log of the inverse probability [18].

Taking an array of terms as input, the R package named texmineR (version 3.0.4) obtains using the TermDocFreq function [48] a data matrix with columns for term frequency, document frequency, and weighted inverse document frequency [49]. This package allows applying lemmatization and elimination of those words that the researcher considers 'noisy'—such as adjectives, articles, or other such words commonly called 'stop words'.

2.3. Latent Dirichlet Assignment (LDA)

The latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus, the basic idea being that a random mix of latent topics is represented, where each topic is characterized by a distribution over words [20]. LDA is a Bayesian variant of probabilistic latent semantic analysis (PLSA), whose predecessor is latent semantic analysis (LSA), LDA is based on a set of assumptions, which states that words in a text are interchangeable between documents and that documents they are represented as a string of individual words that make up the document.

The probabilistic generative process is defined by Blei and Lafferty [50] and is represented in Figure 2. To give a better understanding of the LDA graph, the observed data are the words of a document, and the hidden variables represent the structure of the latent topics. The interaction between the observed documents and the structure of the topics is manifested in the generative process associated with LDA. The generative process will be rewritten, but only the steps for generating the entire document collection will be presented.

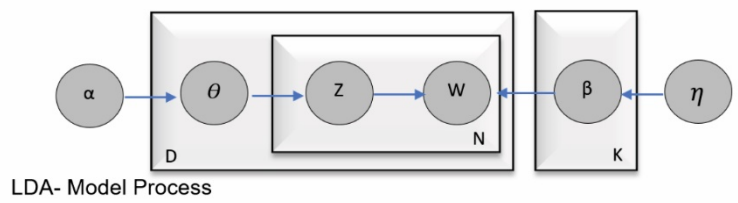

Figure 2. Probabilistic generative process LDA.

The LDA generative process assumes the documents come, is described as:

1. For each topic K
   - I. Draw a distribution over the words (i.e., vocabulary V), $\beta_k \sim Dir_V(\eta)$
2. For each document D:
   - I. Draw a distribution over topics (i.e., ratio of the topic to document) $\theta_d \sim Dir(\alpha)$
   - II. For each word w within document D:
     - i. Draw a topic assignment, $z_{d,n} \sim \text{Mult}(\theta_d)$ (i.e., topic assignment per word)
     - ii. draw a word $w_{d,n} \sim \text{Mult}\left(\beta_{z_{d,n}}\right)$

Where each topic k comes from a Dirichlet distribution $\beta_k \sim Dir_V(\eta)$ and is a multinomial distribution over the vocabulary, each document D is represented as a distribution of topics and originates from a $\theta_d \sim Dir(\alpha)$. The Dirichlet parameter η defines the probability of words within topics, and α the probability of topics within documents. The joint distribution of all hidden variables β_k, θ_d (document topic ratios within D), $z_{d,n}$ (word topic assignments), and observed variables $w_{d,n}$ (words in documents), is described in Equation (2):

$$P(\beta_k, \theta_D, Z_D, W_D) = \prod_{k=1}^{K} P(\beta_k|\eta) \prod_{d=1}^{D} P(\theta_d|\alpha) \prod_{n=1}^{N} P(Z_{d,n}|\theta d)\, P(W_{d,n}, \beta_K) \quad (2)$$

In LDA, all documents share the same set of topics, but each document shows the corresponding topic in different proportions. This process, analyzed in-depth by Blei [20], is computationally intractable but it is the key to LDA, so approximation methods must be applied both for quantitative calculations; such as the generalization of predictions and documents, and for exploratory tasks. To obtain the LDA model with the data under study, the R textmineR package (version 3.0.4) was used, through the FitLdaModel function, which allows us to fit a Dirichlet latent assignment topic model through Gibbs sampling, which is used to obtain the variational inference approach [51]. This model is composed of three matrices:

theta (θ): distribution of topics on the documents;

phi (Φ): distribution of words on topics;

Gamma (γ): distribution of topics by words;

Equation (3) describes the calculation of the posterior probabilities of the distributions of the observations of seeing the corpus observed in each one of the topics.

$$P(\beta_k,\ \theta_D,\ Z_D \mid W_D) = \frac{P(\ \beta_k,\ \theta_D,\ Z_D,\ W_D)}{P(W_D)} \tag{3}$$

Gibbs sampling provides direct estimates of the topic assignment Z for each word, from the estimates θ is described in Equation (4) and Φ is described in Equation (5) of the within-document topic distributions and word-topic distributions, respectively.

where:

C^{WK} is a matrix of dimension W(words) $\times$ K(topics), where C_{ij}^{WK} is the number of times word i is assigned to topic j.

C^{DK} is a matrix of dimension D (Documents) $\times$ K, where C_{dj}^{DK} is the number of times topic j is assigned to some keyword in document d.

α y β are hyperparameters, which act as constraints on the model.

$$\theta'^{(d)}_{j} = \frac{C_{dj}^{DK} + \alpha}{\sum_{k=1}^{K} C_{dk}^{DK} + K \propto} \tag{4}$$

$$\Phi'^{(j)}_{i} = \frac{C_{ij}^{WK} + \beta}{\sum_{k=1}^{W} C_{kj}^{WK} + W\beta} \tag{5}$$

Both expressions according to the interpretations of Steyvers and Griffiths [51] correspond to the predictive distributions of sampling a new i-th word from the j-th topic/topic and sampling a new word (not yet observed) in document d from the j-th topic/topic. Once the model has been created, its goodness of fit is evaluated using the well-known coefficient of determination R^2 applied to topic models. This figure of merit is interpreted in the usual way, as the proportion of variability in the data explained by the model [52]. The textmineR package also allows calculating several indicators on the LDA model, such as the probabilistic coherence [53], which can be interpreted as an estimate of the comprehensibility of a topic by a human.

2.4. HJ-Biplot

Biplots, proposed by Gabriel [54], are graphical representations of multivariate data, allowing the visualization of three or more variables, like a scatter plot showing the joint distribution of two variables. The HJ-biplot is a multidimensional data technique proposed as an alternative to improve the classical biplots introduced by Gabriel, the GH-biplot achieves a high-quality representation of the variables (column marker), while the JK-biplot achieves a high-quality in the ranges of the individuals (row marker). An alternative to optimize biplot methods described by Galindo [25] proposed a multivariate technique called HJ-biplot.

The HJ-biplot [25] is a multivariate graphical representation of a matrix X using markers $j_1, \ldots, j_f$ for its rows and $h_1, \ldots, h_c$ for its columns, chosen so that both markers can be superimposed in the same reference system with maximum representation quality. It is an evolution of the biplots formulated by Gabriel [54] and both are based on the decomposition singular values [55] of the starting matrix X and the subsequent definition of a lower rank approximation for the same [56]. This representation is described in Equation (6), where the HJ-biplot [25] is defined as

$$X = UDX^T \begin{array}{l} J = U\ D \\ H = V\ D \end{array} \tag{6}$$

where X is the data matrix, U is the matrix orthogonal of data columns containing the eigenvectors of XX^T, V is the matrix orthogonal of data whose columns contain the eigenvectors of $X^T X$, and D is the diagonal matrix containing the eigenvalues of X. The row markers are matched to the rows of the JK biplot markers ($J = UD$); in turn, the column markers of the HJ biplot match the marked columns of the GH biplot ($H = VD$), considering the matrix x centered.

In our study, the HJ-biplot has been applied for the graphical representation of the theta and phi data matrix obtained from the LDA analysis and thus visualize the distribution of the words w, with the Topics K and with the documents D. For the analysis of the data of the present study, the R Package GGBiplotGUI 1.0.9 was used [57], which allows different types of biplot representations to be made through a graphical interface. This package has been used in different publications for the analysis of different types of data—including environmental, genetic, and agronomic data [58].

2.5. LDAvis and Heatmap

To obtain an overview of the topics and the differences between them, as well as to facilitate a graphic review of the words most associated with each topic individually, an alternative used is LDAvis, which is an interactive web-based visualization of the estimated topics by the latent Dirichlet assignment which is created by a combination of R and D3 using the popular D3 JavaScript library [59].

A heatmap is a graphic representation of data where the individual values contained in a matrix are represented as colors, both static and interactive; normally, the rows and columns are reordered by the averages obtained, or according to the restrictions imposed by the user of the package of r. The function is provided natively in R. It produces a high-quality matrix and offers statistical tools to normalize the input data, run clustering algorithms, and visualize the result with dendrograms.

Both representation methods are available in R packages. For the application of the LDAvis method, the LDAvis package (version 0.3.2) and the ComplexHeatmap package (version 2.12.0) were used to compare the results with the obtained in the HJ-biplot.

3. Results

A web scraping technique has been applied, using the Rvest de R package, to the pages dedicated to the coronavirus in the three newspapers understudy and published in the following URLs that correspond to the headlines of the news related to the coronavirus: COVID-19 in Spain, published from 1 January 2019 to 27 May 2022: 'https://elpais.com/noticias/coronavirus/', 'https://www.20minutos.es/busqueda/1/?q=covid+coronavirus/', 'https://www.elmundo.es/e/co/coronavirus.html' (accessed on 27 May 2022).

With the collected data, a matrix of 3 columns and 48,112 rows was built, the first column contains an identifier of each document, the second column contains the name of the newspapers, and the third is the headlines of the news extracted from the website of each newspaper. Table 1 shows the number of web news headlines obtained by each newspaper.

Table 1. Number of headlines for each newspaper.

ID	Newspapers	Frequency
1	El-Pais	19,375
2	El-Mundo	18,547
3	20 M	10,190

The CreateDtm function, from the textmineR package, was applied to create the matrix of document terms (news headlines for each newspaper). Stop words usual in Spanish, among which the word 'Coronavirus' was also included, since it is considered that it would generate noise in the headlines, due to its possibly high frequency of appearance, and punctuation marks were also removed, to separate the words from the titles. The TermDocFreq function was applied to this matrix, obtaining a dgCMatrix DTM with

247,137 words in the columns and 48,112 documents in the rows. Using the package functions, a term frequency matrix is generated with the respective IDF weight of each term. To facilitate a graphical analysis in the present study, cleaning of the words whose frequency is less than 900 repetitions or appearances in the entire corpus of the generated text is carried out, finally obtaining a matrix with only 22 terms of the initially generated corpus. Figure 3 shows part of the most used words to create the LDA model.

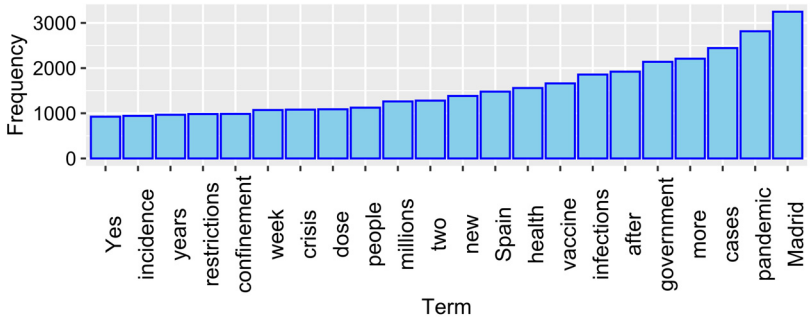

Figure 3. Terms with a higher frequency of matrix DTM.

To generate the LDA model with the FitLdaModel function of the TextmineR package, the optimal number of K topics was first determined, according to the coherence explained by the terms found in each topic. When analyzing the coherence through Figure 4, it is found that 14 are the topics that represent the greatest coherence of the model of 20 possible topics initially evaluated. With the value obtained, an LDA model is generated, restricting it to 14 topics, to obtain the theta Θ, phi Φ, and gamma γ matrices.

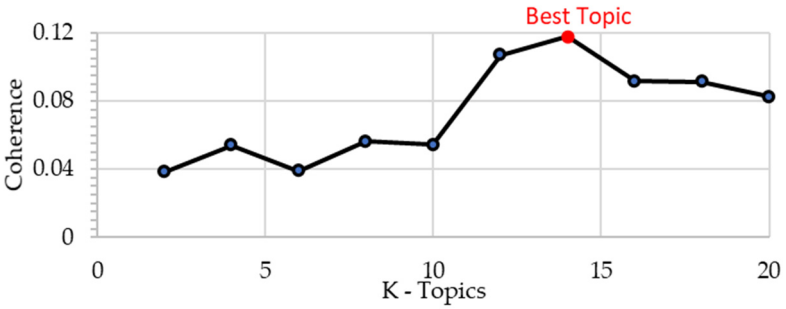

Figure 4. Coherence of topics.

3.1. Results LDAvis and Heatmap

For the analysis of the results obtained in the LDA, it is also represented by means of a heatmap which represents the possibility that a topic K belongs to a newspaper by means of a heatmap, as shown in Figure 5. It is observed as topics t_9 (USA_Pandemic) and t_6 (Vaccines_people) present a higher average frequency in the 20 M newspaper, the topic t_9 also shows to be relevant for the newspaper El Mundo, while in El País it is shown that almost all the topics present a proportion in this figure, the calculation of the averages of the documents of each newspaper with the topics was used, and these values are those represented in Figure 5.

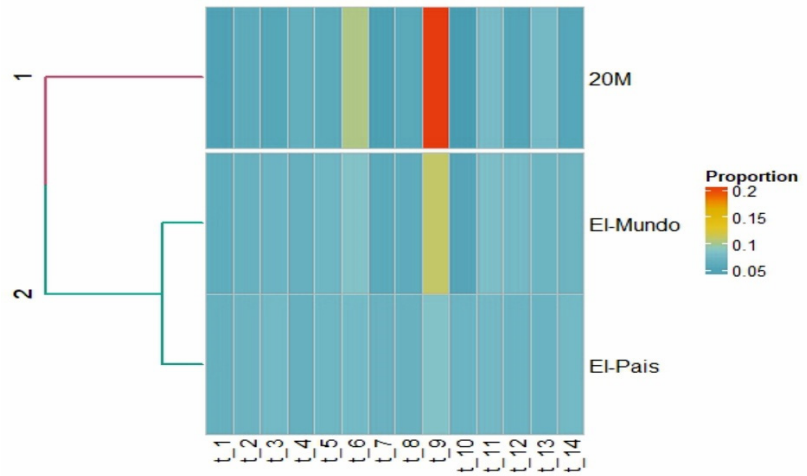

Figure 5. Heatmap of the obtained LDA.

LDAvis allows the visualization of the topics obtained from the Model LDA generated in the study on a web page using an interactive graphic (SVG, scalable vector graphics). With this method, the topics are represented with a circle, the greater the diameter of the circle, the greater the proportion of words in this topic of the model. These circles are represented positioned in a multidimensional scaling plane (MDS), where the distance map between topics is a visualization of these in a two-dimensional space, where the circles are plotted using a multidimensional scaling algorithm based on the words that they contain, so the closer topics have more words in common. The web application allows you to interact with the graph so that when you select the topic you can see the words that make up the selected topic, ordered in decreasing order according to the frequency with which they appear in each topic. As an example of this functionality, Figure 6 shows topic 2 the words "vaccine", "dose", "years", "third", and "older" are positioned among the most representative of the generated topic.

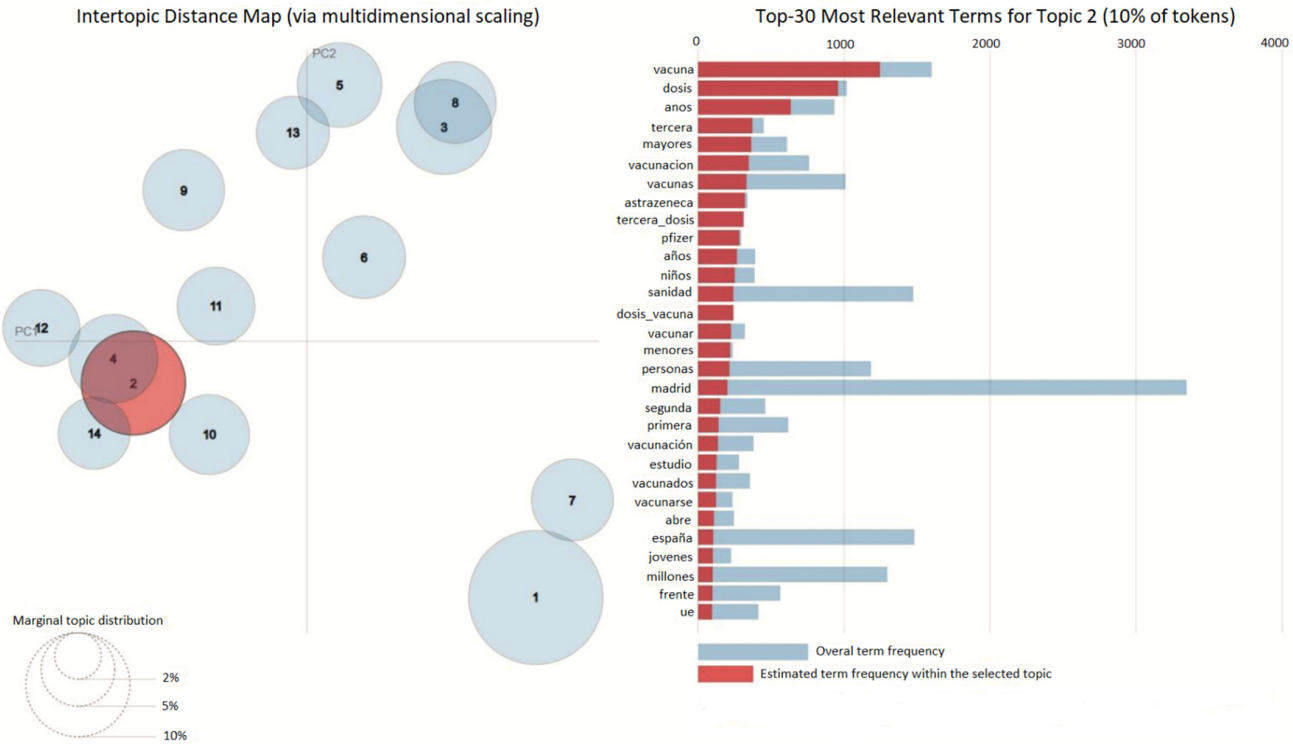

Figure 6. LDAvis representation of topics.

3.2. Results HJ-Biplot

To obtain the HJ-biplot representation of the LDA model obtained, the three matrices θ, Φ, and γ were generated using the summary function of the Base package of r, in which the 14 topics are characterized, obtaining the theta matrices θ (48,112 Documents $\times$ 14 Topics), phi Φ (226 words $\times$ 14 topics), and gamma γ (14 topics $\times$ 226 words). Appendix A shows the matrix of topics K with the terms w, from the corpus of documents D, with the coherence explained by each topic in the model.

For the representation of the topics in the newspapers analyzed by means of the HJ-biplot, a matrix is generated from the matrix θ, obtained in the LDA, since it classifies the topics according to the probability that each of these belongs to each analyzed document. The transposed matrix of theta is generated, to this matrix θ^T, the average of each document $\overline{d}$ is calculated (this equation applies from i document to n document for each newspaper), which in this matrix represents the possibility that each document d belongs to each topic K, the average is calculated for each set of newspapers, D, thus obtaining the possibility that each topic K belongs to the respective newspapers analyzed, this process is described in Figure 7 and Equation (7). That is a representation of the topics K is made for each one of the documents D (in this case D being the newspapers), forming a new matrix X.

$$\overline{d}_i = \frac{d_{i1} + d_{i2} \ldots + d_{in}}{D} \tag{7}$$

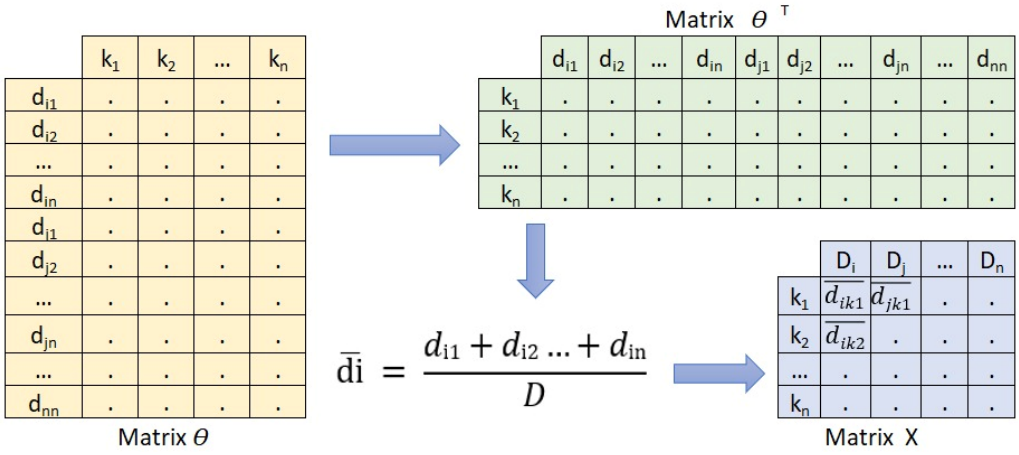

Figure 7. Process of obtaining the X matrix for HJ-biplot.

The interpretation of the HJ-biplot obtained in the present study is shown in Figure 8. Rules like those used in multidimensional scaling (MDS), correspondence analysis, factor analysis, and classical biplot are used. Thus, the length of the markers corresponding to the columns (vectors) approximate the standard deviation of the topics; the cosines of the angles between the markers (vectors) column approximate the correlations between news newspapers. To understand HJ-biplot, let us consider the order of the orthogonal projections of the row markers (points) onto a column marker (vector) approximates the order of the row elements (centers) in that column (the same property holds for the projection of the markers column in the direction defined by a row marker). Acute angles are associated with newspapers with a positive correlation (20 M and El Mundo), whereas obtuse angles indicate negative correlation and right angles indicate variables unrelated (20 M and El País, for example). Likewise, the cosines of the angles between the topic markers and the axes (principal components) approximate the correlations between the two. The greater the projection of a point on a vector, the more the center deviates from the mean of the daily news. The distances among row markers are interpreted as an inverse function of their similarities, in such a way that closer markers (topics) are more similar.

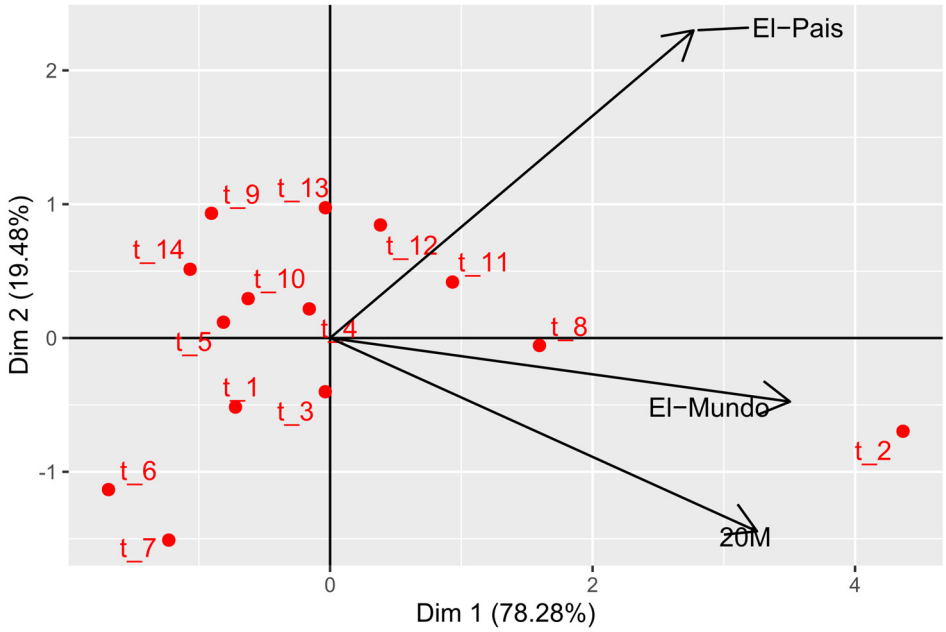

Figure 8. HJ-biplot of topics obtained in LDA.

In the first place, it is observed that—overall—there is a good quality of representation in this first factorial plane, with 97.76% of inertia absorbed or explained. In Table 2, the names of the 14 topics obtained in the LDA model are observed. The topic t_12 "Government_Madrid" is explained in the headlines of the newspaper El País since the projection of this topic on the vector that represents the newspaper "El País" is much higher than the projection of the same topic. The other two vectors/markers represent El Mundo and 20M, in which this projection is almost null. Although the most explained topic in El Mundo would correspond to t_2 "New_Contagions" which also has an explanation in the other two newspapers, although less than in the first. Topics t_8 "Dose_Vaccine" and t_11 "Curfew" are mainly related to the newspapers El País, El Mundo, and 20M. The relationship between newspapers is also observed in this representation. According to the topics covered in the headlines related to COVID, El Mundo and 20 Minutos are perceived as having a high relationship with each other in the topics covered since the angle between the vectors that represent both of markers is smaller. However, El País presents other topics that it delves into in greater detail, since it does not present a good correlation with the other two newspapers, as it presents an angle of practically 90° between the vectors. It is observed how certain topics are related only to some news newspapers; but in general, the three have most of the most similar topics in their publications.

Table 2. Label topics.

Topics	Label Topics	Topics	Label Topics
t_1	Positive_Cases	t_8	Dose_Vaccine
t_2	New_Contagions	t_9	USA_Pandemic
t_3	New_Variant	t_10	Wave_Pandemic
t_4	Third_dose	t_11	Curfew
t_5	Community_Madrid	t_12	Government_Madrid
t_6	Vaccines_people	t_13	Government_Millions
t_7	Measures_Pedro_Sanchez	t_14	House_confinement

4. Discussion

The LDAvis represents the topics in a two-dimensional space without considering the periodicals; in addition, this package visualizes the topics as circles in the two-dimensional plane whose centers are determined by calculating the Jensen–Shannon divergence [60]

between the topics and then using a multidimensional scale to project the distances between subjects in two dimensions. The overall prevalence of each theme is coded using the areas of the circles. Therefore, the results differ from the LDA obtained when executing the FitLdaModel function of the TextminerR package, where the Gibbs sampling is considered, as well as the values given to the hyperparameters for obtaining the topics and the grouping of the words according to the given weight. By the frequency in each topic, this can also be controlled by the number of iterations assigned to obtain the model.

In the representation of the LDA model obtained by means of a heatmap, the averages of the possibility that each document of a newspaper belongs to a certain topic are observed in a grouped manner, these average values obtained by each newspaper are shown as the possibility that each topical has a higher value of the frequency of belonging to a particular newspaper. Additionally, it can be parameterized so that the newspapers are displayed in a grouped manner, the distance considered in this method is the Euclidean [61], and in this representation, the row markers (topics) are given greater representativeness.

Biplot techniques are based on the same principles on which most dimensionality reduction factorial techniques are based. The fundamental difference is that a joint representation of rows and columns is incorporated, unlike principal component analysis which reduces the column data to a smaller number of components seeking to explain as much of the total variance in the variables as possible by calculating the components as linear combinations of the original variables [62], or unlike analysis factorial of correspondence that is used when there is a significant association between the categorical variables studied, representing the rows and columns of the contingency table in two reduced vector spaces, to later superimpose them and obtain the joint representation of both [63].

The HJ-biplot method, which is presented as an alternative for the representation of the results obtained in an LDA content modeling, can obtain a high-quality representation simultaneously in the row markers (topics) and column markers (newspapers), enabling the study of the correlation between documents and visualizing the topics according to the corpus of the document with which it has the greatest representativeness. The distance between the row markers (topics) enables the identification of clusters of individuals with similar profiles. Any hierarchical or non-hierarchical clustering technique can be used to help identify relevant clusters.

5. Conclusions

The topic model is an unsupervised method applied to text mining. In this study, it was applied to news from digital newspapers, where the HJ-biplot is presented as a new option to visualize newspapers and topics with the highest quality of representation, which is not possible with other traditional biplot models. Two comparative methods of representation of the LDA model were used: the heatmap represents the topics with better quality and the LDAvis does not consider the newspapers within its multidimensional scaling representation, which does not allow exploration of the possible relationships that the topic has with the newspapers. Therefore, it is not possible to observe which topic contributes to a newspaper, unlike with the HJ-biplot and the heatmap which allow this analysis, but in a different way between them.

It is recommended that the effect of applying different methods of selecting and extracting data from the web be explored, as well as applying other methods associated with LDA to obtain the topics, and even applying machine learning methods for the representation in the HJ-biplot of the topics and digital newspapers.

Author Contributions: Conceptualization, L.P.-B., P.G.-V. and F.D.-Á.; methodology, L.P.-B.; validation, L.P.-B., P.G.-V. and F.D.-Á.; formal analysis, L.P.-B.; investigation, L.P.-B.; data curation, L.P.-B.; writing—original draft preparation, L.P.-B.; writing—review and editing, L.P.-B., P.G.-V. and F.D.-Á.; supervision, P.G.-V. and F.D.-Á. 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 required.

Informed Consent Statement: Not required.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/Pilacuan-Bonete-Luis/Data_HJ-Biplot_newspapers/blob/main/Data_newspapers.xlsx (accessed on 27 May 2022).

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

Appendix A

In Table A1: each topic is observed with its respective group of words obtained in the LDA model, as well as the coherence which gives us an idea of how coherent a model is in terms of the distribution of its topics. The more different the words of the topics are among themselves, the less related the topics will be and the better coherence the model will have. As for prevalence, which is a measure of how sensitive the model is when data that are not part of it are added before, the lower the prevalence value is the better the model is.

Table A1. Topics, label topics, coherence, prevalence, and top term of the model LDA.

Topics	Label Topics	Coherence	Prevalence	Top Terms
t_1	Positive_Cases	0.046	6.533	positive, gives, dies, months, years, days, three, quarantine, give, after, hospital, test, first, four, Madrid, six, five, UCI, pandemic, days, home, ago, case, life, trump, virus, anus, seven, fear, years, weeks, people, health, Barcelona, mask, patients, USA, returns, PCR, less
t_2	New_Contagions	0.138	11.359	cases, new, infections, incidence, health, deaths, registered, deceased, hours, Spain, new_cases, week, dead, positive, sum, notifies, UCI, Balearic, islands, low, new_infections, day, Madrid, last, figure, last_hours, continues, exceeds, Cantabria, Euskadi, pandemic, almost, hospitalized, last, income, increase, four, Spain, risk, three, decrease
t_3	New_Variant	0.166	7.178	new, variant, omicron, kingdom, united, united_kingdom, span, Europe, virus, normality, who, quarantine, Johnson, first, Spain, case, restrictions, confinement, pandemic, country, vaccinated, asi, Germany, alert, risk, puts, EU, vaccines, France, Italy, new, test, vaccine, wave, cases, return, tourism, united, returns, experts
t_4	Third_dose	0.021	6.919	residences, Madrid, hospital, outbreak, health, residence, elderly, hospitals, elderly, health, patients, centers, infected, test, positive, people, health, three, new, workers, UCI, virus, Valencia, masks, missing, PCR, pandemic, four, Generalitat, less, dead, leave, health, board, Catalonia, Barcelona, get vaccinated, Ayuso, ask, death
t_5	Community_Madrid	0.205	6.359	community, Madrid, Valenciano, valencia_community, madrid_community, phase, Monday, test, de-escalation, health, week, closure, Generalitat, leave, restrictions, cvirus, government, passport, vaccination, Catalonia, Ayuso, health, residences, PCR, may, masks, Barcelona, requests, infections, measures, people, mask, pandemic, centers, vaccination, hospitals, bars, areas, leisure, Spain

Table A1. *Cont.*

Topics	Label Topics	Coherence	Prevalence	Top Terms
t_6	Vaccines_people	0.013	5.8	vaccines, people, pandemic, millions, Spain, complete, vaccination, world, population, like this, year, less, rioja, almost, front, half, first, exceeds, vaccine, greater, EU, Galicia, ago, country, Europe, months, Spain, study, virus, six, crisis, summer, three, great, children, dose, USA, vaccinated, years, five
t_7	Measures_Pedro_Sanchez	0.146	6.272	Sanchez, cases, active, Pedro, active_cases, Pedro_Sanchez, day, Galicia, rioja, infections, alarm, days, deceased, decrease, increase, municipalities, new, new, Sanchez, government, positive, new_contagions, week, pp, hospitalized, Spain, requests, exceeds, less, announces, citizens, low, cvirus, crisis, continues, centers, front, maintains, pandemic, plan
t_8	Dose_Vaccine	0.121	8.65	vaccine, dose, years, third, older, vaccination, vaccines, AstraZeneca, third_dose, Pfizer, years, children, health, vaccine_dose, vaccinate, minors, people, Madrid, second, first, vaccination, study, vaccinated, get vaccinated, open, Spain, young people, millions, front, EU, population, residences, week, risk, less, USA, leave, health, experts, months
t_9	USA_Pandemic	0.114	6.2	pandemic, China, Mexico, USA, USA, uu, city, who, world, trump, vaccine, virus, united, new, new, alert, crisis, vaccination, first, great, vaccines, Johnson, year, millions, Europe, case, USA, how, outbreak, health, variant, major, ask, medium, announce, normality, home, health, government, EU
t_10	Wave_Pandemic	0.066	6.51	pandemic, wave, first, second, time, risk, Spain, virus, greater, worse, Spain, contagion, infections, world, middle, year, ICU, life, crisis, death, third, year, fear, new, confinement, new, alert, low, Europe, less, incidence, front, hospitals, asi, month, study, week, health, health, China
t_11	Curfew	0.276	7.89	restrictions, Madrid, curfew, touch, touch, stay, measures, closure, new, confinement, people, government, Catalonia, bars, leisure, request, municipalities, passport, areas, infections, Barcelona, week, Catalonia, Andalusia, Christmas, close, alarm, health, communities, mask, curb, Generalitat, meeting, avoid, start, Monday, three, healthcare, weigh, six, maintain
t_12	Government_Madrid	0.027	7.316	government, Madrid, Ayuso, ask, pp, crisis, measures, alarm, masks, pandemic, Sanchez, citizens, plan, front, ask, Sanchez, confinement, communities, de-escalation, health, avoid, test, says, new, virus, now, health, mask, EU, phase, lack, residences, sanitary, return, weigh, announce, Christmas, sanitary, put, data

Table A1. *Cont.*

Topics	Label Topics	Coherence	Prevalence	Top Terms
t_13	Government_Millions	0.023	6.925	millions, government, pandemic, euros, crisis, companies, erte, aid, workers, tourism, sector, plan, Spain, masks, Barcelona, year, summer, less, announce, almost, request, Generalitat, Madrid, year, test, health, board, work, month, middle, half, front, first, may, EU, health, new, Sanchez, leave, Spain
t_14	House_confinement	0.004	6.09	home, confinement, Madrid, quarantine, children, mask, how, students, day, homecoming, course, street, masks, alarm, first, today, Spain, work, pandemic, so, Monday, leave, may, government, less, Catalonia, Barcelona, de-escalation, now, children, days, phase, restrictions, life, avoid, fear, France, close, day, normality

References

1. He, W.; Zha, S.; Li, L. Social Media Competitive Analysis and Text Mining: A Case Study in the Pizza Industry. *Int. J. Inf. Manag.* **2013**, *33*, 464–472. [CrossRef]
2. Alalwan, A.A.; Rana, N.P.; Dwivedi, Y.K.; Algharabat, R. Social Media in Marketing: A Review and Analysis of the Existing Literature. *Telemat. Inform.* **2017**, *34*, 1177–1190. [CrossRef]
3. Pejic-Bach, M.; Bertoncel, T.; Meško, M.; Krstić, Ž. Text Mining of Industry 4.0 Job Advertisements. *Int. J. Inf. Manag.* **2020**, *50*, 416–431. [CrossRef]
4. De la Hoz-M, J.; Fernández-Gómez, M.J.; Mendes, S. LDAShiny: An R Package for Exploratory Review of Scientific Literature Based on a Bayesian Probabilistic Model and Machine Learning Tools. *Mathematics* **2021**, *9*, 1671. [CrossRef]
5. Slobodin, O.; Plochotnikov, I.; Cohen, I.-C.; Elyashar, A.; Cohen, O.; Puzis, R. Global and Local Trends Affecting the Experience of US and UK Healthcare Professionals during COVID-19: Twitter Text Analysis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 6895. [CrossRef]
6. WHO. *COVID-19 Weekly Epidemiological Update*; WHO: Geneva, Switzerland, 2022.
7. Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **2020**, *382*, 727–733. [CrossRef]
8. Brüssow, H. The Novel Coronavirus—A Snapshot of Current Knowledge. *Microb. Biotechnol.* **2020**, *13*, 607–612. [CrossRef]
9. McKibbin, W.J.; Fernando, R. The Global Macroeconomic Impacts of COVID-19: Seven Scenarios. *SSRN Electron. J.* **2020**, *20*, 1–30. [CrossRef]
10. 20Minutos. ¿Cuál Fue El Primer Caso de Coronavirus en España y en La Península? Available online: https://www.20minutos.es/noticia/4186871/0/coronavirus-primer-caso-espana-peninsula/ (accessed on 15 April 2020).
11. Estudio General de Medios Ranking de Diarios. Available online: http://reporting.aimc.es/index.html#/main/diarios (accessed on 16 April 2020).
12. Mutz, D.C.; Soss, J. Reading Public Opinion: The Influence of News Coverage on Perceptions of Public Sentiment. *Public Opin. Q.* **1997**, *61*, 431. [CrossRef]
13. Hoffman, L.H.; Glynn, C.J.; Huge, M.E.; Sietman, R.B.; Thomson, T. The Role of Communication in Public Opinion Processes: Understanding the Impacts of Intrapersonal, Media, and Social Filters. *Int. J. Public Opin. Res.* **2007**, *19*, 287–312. [CrossRef]
14. Peretti, P.O.; Lucas, C. Newspaper Advertising Influences on Consumers' Behavior by Socioeconomic Status of Customers. *Psychol. Rep.* **1975**, *37*, 693–694. [CrossRef]
15. Thornton, J.A.; Wahl, O.F. Impact of a Newspaper Article on Attitudes toward Mental Illness. *J. Community Psychol.* **1996**, *24*, 17–25. [CrossRef]
16. Baumgartner, R.; Gatterbauer, W.; Gottlob, G. Web Data Extraction System. *Encycl. Database Syst.* **2009**, *1*, 3465–3471.
17. Collobert, R.; Weston, J.; Com, J.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537. [CrossRef]
18. Jones, K.S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *J. Doc.* **1972**, *28*, 11–21. [CrossRef]
19. Deerwester, S.; Harshman, R.; Susan, T.; George, W.; Thomas, K. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
20. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022. [CrossRef]
21. Aldjanabi, W.; Dahou, A.; Al-Qaness, M.A.A.; Elaziz, M.A.; Helmi, A.M.; Damaševičius, R. Arabic Offensive and Hate Speech Detection Using a Cross-Corpora Multi-Task Learning Model. *Informatics* **2021**, *8*, 69. [CrossRef]
22. Hadwan, M.; Al-Sarem, M.; Saeed, F.; Al-Hagery, M.A. An Improved Sentiment Classification Approach for Measuring User Satisfaction toward Governmental Services' Mobile Apps Using Machine Learning Methods with Feature Engineering and SMOTE Technique. *Appl. Sci.* **2022**, *12*, 5547. [CrossRef]

23. Sievert, C.; Shirley, K.E. LDAvis: A Method for Visualizing and Interpreting Topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, MD, USA, 27 June 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 63–70.
24. Raivo Kolde. cran.r-project.org/package=pheatmap— Pheatmap: Pretty heatmaps. R Package Heatmap version 2.12.00. 2019. Available online: https://cran.r-project.org/web/packages/pheatmap/index.html/ (accessed on 15 April 2022).
25. Galindo-Villardón, P. Una Alternativa de Representacion Simultánea: HJ-Biplot (An Alternative of Simultaneous Representation: HJ-Biplot). *Questío* **1986**, *10*, 13–23.
26. Díaz-Faes, A.A.; González-Albo, B.; Galindo, M.P.; Bordons, M. HJ-Biplot Como Herramienta de Inspección de Matrices de Datos Bibliométricos. *Revista Española Documentación Científica* **2013**, *36*, e001. [CrossRef]
27. Julia, D.C.; Galindo, P.V.; Villardón, M.P.G. Grupos de Discusión y HJ-Biplot: Una Nueva Forma de Análisis Textual. *Revista Ibérica Sistemas Tecnologias Informação* **2014**, *E2*, 19–35. [CrossRef]
28. Zulaima, O.M. *Contribuciones al Análisis de Datos Textuales*; Universidad de Salamanca: Salamanca, Spain, 2006.
29. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 3rd ed.; Elsevier Inc.: Amsterdam, The Netherlands, 2012; ISBN 9780123814791.
30. Fayyad, U.; Stolorz, P. Data Mining and KDD: Promise and Challenges. *Futur. Gener. Comput. Syst.* **1997**, *13*, 99–115. [CrossRef]
31. Alyahyan, E.; Düştegör, D. Predicting Academic Success in Higher Education: Literature Review and Best Practices. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 3. [CrossRef]
32. The R Foundation R 2020. Available online: https://www.r-project.org/ (accessed on 1 May 2021).
33. Manning, C.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008; ISBN 9780511809071.
34. Markov, Z.; Larose, D.T. *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*; John Wiley & Sons: Hoboken, NJ, USA, 2007; ISBN 0470108088.
35. Kamath, S.S.; Bagalkotkar, A.; Khandelwal, A.; Pandey, S.; Poornima, K. Sentiment Analysis Based Approaches for Understanding User Context in Web Content. In Proceedings of the 2013 International Conference on Communication Systems and Network Technologies, CSNT 2013, Gwalior, India, 6–8 April 2013; pp. 607–611.
36. Catanese, S.A.; De Meo, P.; Ferrara, E.; Fiumara, G.; Provetti, A. Crawling Facebook for Social Network Analysis Purposes. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics, Sogndal, Norway, 25–27 May 2011; ACM Press: New York, NY, USA, 2011; p. 1.
37. Chandler, J.D.; Salvador, R.; Kim, Y. Language, Brand and Speech Acts on Twitter. *J. Prod. Brand Manag.* **2018**, *27*, 375–384. [CrossRef]
38. Plake, C.; Schiemann, T.; Pankalla, M.; Hakenberg, J.; Leser, U. ALIBABA: PubMed as a Graph. *Bioinformatics* **2006**, *22*, 2444–2445. [CrossRef]
39. Xie, X.; Fu, Y.; Jin, H.; Zhao, Y.; Cao, W. A Novel Text Mining Approach for Scholar Information Extraction from Web Content in Chinese. *Futur. Gener. Comput. Syst.* **2019**, *111*, 859–872. [CrossRef]
40. Schedlbauer, J.; Raptis, G.; Ludwig, B. Medical Informatics Labor Market Analysis Using Web Crawling, Web Scraping, and Text Mining. *Int. J. Med. Inform.* **2021**, *150*, 104453. [CrossRef]
41. Rekik, R.; Kallel, I.; Casillas, J.; Alimi, A.M. Assessing Web Sites Quality: A Systematic Literature Review by Text and Association Rules Mining. *Int. J. Inf. Manag.* **2018**, *38*, 201–216. [CrossRef]
42. Zhao, Y. *R and Data Mining: Examples and Case Studies*; Academic Press: Cambridge, MA, USA; Elsevier: Amsterdam, The Netherlands, 2012; ISBN 9780123969637.
43. Khalil, S.; Fakir, M. RCrawler: An R Package for Parallel Web Crawling and Scraping. *SoftwareX* **2017**, *6*, 98–106. [CrossRef]
44. Wickham Hadley Easily Harvest (Scrape) Web Pages 2019. Available online: https://rvest.tidyverse.org/ (accessed on 1 May 2021).
45. Salton, G.; Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
46. Aizawa, A. An Information-Theoretic Perspective of Tf-Idf Measures. *Inf. Process. Manag.* **2003**, *39*, 45–65. [CrossRef]
47. Luhn, H.P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM J. Res. Dev.* **1957**, *1*, 309–317. [CrossRef]
48. Thomas, J. Función TermDocFreq | RDocumentation 2019. Available online: https://www.rdocumentation.org/packages/textmineR/versions/3.0.4/topics/TermDocFreq (accessed on 1 May 2021).
49. Tommy, J.; William, D. Functions for Text Mining and Topic Modeling 2019. Available online: https://www.rtextminer.com/ (accessed on 1 May 2021).
50. Blei, D.M.; Lafferty, J.D. Topic Models. In *Text Mining: Classification, Clustering, and Applications*; Taylor & Francis Group, Ed.; Chapman and Hall/CRC: New York, NY, USA, 2009; pp. 71–82. ISBN 9780429191985.
51. Steyvers, M.; Griffiths, T. Probabilistic Topic Models. In *Handbook of Latent Semantic Analysis*; Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W., Eds.; Laurence Erlbaum: Mahwah, NJ, USA, 2006; pp. 427–448. ISBN 1135603286.
52. Jones, T. A Coefficient of Determination for Probabilistic Topic Models. *arXiv* **2019**, arXiv:1911.11061. [CrossRef]
53. Rosner, F.; Hinneburg, A.; Röder, M.; Nettling, M.; Both, A. Evaluating Topic Coherence Measures. *arXiv* **2014**, arXiv:1403.6397. [CrossRef]

54. Gabriel, K.R. The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika* **1971**, *58*, 453–467. [CrossRef]
55. Eckart, C.; Young, G. The Approximation of One Matrix by Another of Lower Rank. *Psychometrika* **1936**, *1*, 211–218. [CrossRef]
56. Eckart, C.; Young, G. A Principal Axis Transformation for Non-Hermitian Matrices. *Bull. Am. Math. Soc.* **1939**, *45*, 118–121. [CrossRef]
57. Frutos, E.; Galindo, M.P. cran.r-project.org/package=GGEBiplotGUI. GGEBiplotGUI 2016. Available online: https://cran.r-project.org/web/packages/GGEBiplotGUI/index.html (accessed on 1 May 2021).
58. Frutos, E.; Galindo, M.P.; Leiva, V. An Interactive Biplot Implementation in R for Modeling Genotype-by-Environment Interaction. *Stoch. Environ. Res. Risk Assess.* **2014**, *28*, 1629–1641. [CrossRef]
59. Bostock, M.; Ogievetsky, V.; Heer, J. D3 Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 2301–2309. [CrossRef]
60. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
61. Zuguang, G. Packages ComplexHeatmap. 2021. Available online: https://www.bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html (accessed on 1 May 2021).
62. Pearson, K. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *6*, 559–572. [CrossRef]
63. Benzécri, J.-P. *L'analyse Des Données. Tomo I: La Taxonomie*; Dunod: Paris, France, 1973; Volume 2, ISBN 2040071539.