Text mining: An exploratory look at the proposed aspects of improvement in measurement and social development in Mexico: Selection of Academic Researchers for CONEVAL 2020

Medel-Ramírez, Carlos and Medel-López, Hilario

Universidad Veracruzana / Instituto de Investigaciones y Estudios Superiores Económicos y Sociales, Universidad Veracruzana / Instituto de Antropología

15 July 2020

Minería de texto

Una mirada exploratoria a las vertientes propuestas de mejora en la medición y desarrollo social en México: Selección de Investigadores Académicos para el CONEVAL 2020

Text mining

An exploratory look at the proposed aspects of improvement in measurement and social development in Mexico: Selection of Academic Researchers for CONEVAL 2020

Authors

Carlos Medel-Ramírez[1] , Hilario Medel-López[2]

Academic affiliation

1. Universidad Veracruzana / Instituto de Investigaciones y Estudios Superiores Económicos y Sociales
2. Universidad Veracruzana / Instituto de Antropología

Corresponding author

Carlos Medel-Ramírez (cmedel@uv.mx)

Abstract

The importance of the working document is that it allows analyzing the position of 104 researchers of the Conacyt National System of Researchers (Candidate levels, Level I, Level II and Level III) regarding their perspective on modifications and / or improvements in the evaluation and the measurement of poverty for the development of social policy, facing the challenges for the consolidation of evaluation in Mexico. This position arises as part of the prospective exercise in accordance with the Call dated January 9, 2020 regarding the process of electing three academic researchers to form part of the National Council for the Evaluation of Social Development Policy in the National Council for the Evaluation of the Social Development Policy (CONEVAL). The text analysis is carried out by designing a text mining algorithm using the Orange Data Mining 3.26.0 software, in order to identify thematic groups of the positioning, to identify the opportunity area and to serve as the basis for the analysis and design of public policy on Social Development in Mexico.

Keywords

Text mining, Social Development, Indicators, Evaluation, CONEVAL, Poverty Measurement, Evaluation, Mexico

Specifications table

Topic	Desarrollo Social y Evaluación
Specific subject area	Text mining analysis regarding its perspective on modifications and / or improvements in the evaluation and measurement of poverty for the development of social policy, facing the challenges for the consolidation of evaluation in Mexico.
Type of data	Table Figure
Data source	The information comes from the website of the National Council for the Evaluation of Social Development Policy (CONEVAL) https://www.coneval.org.mx/quienessomos/InvestigadoresAcademicos/Paginas/Convocatoria_Eleccion_2020.aspx Files are stored in this repository PDF format containing individualized documents of 104 researchers of the National System of Level Investigators (Candidate, I, II and III) participating in the Call for the election of three academic researchers to form part of the National Council for the Evaluation of Social Development Policy for CONEVAL 2020, dated January 9, 2020. The information corresponds to the "public version" incorporated in the CONEVAL site, which has been marked as reserved those sensitive personal data and that in accordance with article 113, section I of the Federal Law of Transparency and Access to Public Information, it is reserved with a shading made by CONEVAL
Model and instruments used:	Text mining is used for the analysis of 104 documents in PDF format, these documents express the individualized perspective of the importance of the evaluation and measurement of poverty for the development of social policy and the challenges for the consolidation of the evaluation in Mexico. The analysis of the information is carried out using the Orange Data Mining software version 3.26.0, for which the s modules are used: a) text preprocessing module, b) module for the construction of vector spaces, c) module for the word bag integration and topic modeling and d) visualization module as final map of the word cloud. The information analysis uses the Agglomerative Hierarchical Methods to identify clusters that express similarity of thematic opinion of the 104 researchers of the National System of Level Investigators (Candidate, I, II and III) on the importance of evaluation and Poverty measurement for the development of social policy and the challenges for the consolidation of evaluation in Mexico. The Ward Method of the corresponding module is also used in the Orange Data Mining software version 3.26.0, to identify the similarity in thematic aspects, derived from the improvement documents expressed by the 104 SNI researchers, participants in the call for the election of three academic researchers to form part of the National Council for the Evaluation of Social Development Policy (CONEVAL 2020).

Software	Orange Data Mining software version 3.26.0 Orange is free software under the terms of the GNU General Public License published by the Free Software Foundation. Recovered from: https://orange.biolab.si/ OCR.space API version 3.50 Server based OCR software for automatic document capture and PDF conversion. Recovered from: https://ocr.space/
Data format	The information in PDF format has been available since January 28, 2020 on the site https://www.coneval.org.mx/quienessomos/InvestigadoresAcademicos/Paginas/Convocatoria_Eleccion_2020.aspx
Parameters for data collection.	The information is presented in PDF (Portable Document Format), it corresponds to 104 personalized documents, each document consists of at least 4 pages in free format and in Spanish.
Description of data collection.	The information in PDF documents have at least two characteristics: a) Legible for text mining treatment and b) Not legible for text mining treatment. PDF documents whose characteristics are "Not readable for text mining treatment" underwent a process of Optical Character Recognition (OCR) to transform PDF documents into a readable format and prepare them for the use of the various modules of text mining of Orange Data Mining software version 3.26.0.
Data source location	Country: Mexico National Council for the Evaluation of Social Development Policy (CONEVAL 2020)
Data accessibility	Raw data can be retrieved from the Github repository https://github.com/CMedelR/mapatotonacapan/edit/master/README.md

Data value

The text mining algorithm allows to identify similarity in thematic aspects expressed by 104 SNI Conacyt researchers as participants in the call for the election of three academic researchers to form part of the National Council for the Evaluation of Social Development Policy (CONEVAL 2020) , identifying clusters, according to the themes identified and proposed for the improvement of the Social Development Policy, as part of the proposal for action and change to prepare scenarios for making policy decisions on social development in Mexico.

Data description

Be:

SIN = National System of Researchers of CONACYT

Where:

SNI Level = SNI researcher according to category (Candidate, I, II, III)

Be:

$$\text{CONEVAL}^{\text{Academic Researcher Candidate}} = \text{Call for election of CONEVAL 2020 Academic Researcher}$$

So:

$$\text{SNI}^{\text{Level}}\,\text{CONEVAL}^{\text{Academic Researcher Candidate}} = \text{SNI Researcher according to category (Candidate, I, II, III) participating in the Call for the election of CONEVAL 2020 Academic Researcher}$$

donde:

n = Total number of SNI Researchers according to category (Candidate, I, II, III) participating in the Call for the election of CONEVAL 2020 Academic Researcher

n = 104 SNI researchers according to category (Candidate, I, II, III) participating in the Call for the election of the CONEVAL 2020 Academic Researcher

Teniéndose que:

$$\sum_{i=1}^{104} \left(\text{SNI}^{(\text{Level})}_{\text{CONEVAL Academic Researcher Candidate}} \right) = \left[\text{SNI}^{(\text{Candidate})} + \text{SNI}^{(\text{I})} + \text{SNI}^{(\text{II})} + \text{SNI}^{(\text{III})} \right]_{\text{CONEVAL Academic Researcher Candidate}}$$

Where:

$\text{SNI}^{(\text{Candidate})}$ = Researcher of the National System of Researchers with Candidate level
$\text{SNI}^{(\text{I})}$ = Researcher of the National System of Researchers with level I
$\text{SNI}^{(\text{II})}$ = Researcher of the National System of Researchers with level II
$\text{SNI}^{(\text{III})}$ = Researcher of the National System of Researchers with level III

And that as part of the process for the election of CONEVAL 2020 Academic Researchers, in accordance with the second base section A) of the Call for the election of three academic researchers to form part of the National Council for the Evaluation of Social Development Policy dated January 9, 2020, which notes

> "… A document with an autograph signature, of no more than 4 pages written in 12-point arial, single-spaced, stating your willingness to be considered as a candidate to join the National Council for the Evaluation of Social Development Policy, as well as his perspective of the importance of the evaluation and measurement of poverty for the development of social policy and the challenges for the consolidation of evaluation in Mexico… "

So:

$$\sum_{i=1}^{104} [\,(\text{SNI}^{(\text{Level})}{}_{\text{CONEVAL Academic Researcher Candidate}})\,]^{\text{DataTEXT}} = \text{Positioning document in the CONEVAL 2020 Call process}$$

De donde:

$$\sum_{i=1}^{104} [\,(\text{SNI}^{(\text{Nivel})}{}_{\text{CONEVAL Academic Researcher Candidate}})\,]^{\text{DataTEXT}}$$

It seeks to identify the following aspects in:

$$\sum_{i=1}^{104} [\,(\text{SNI}^{(\text{Nivel})}{}_{\text{CONEVAL Academic Researcher Candidate}})\,]^{\text{DataTEXT}}$$

1. Percentage participation of SNI (Level) CONEVAL Academic Researcher Candidate, by sex.
2. Percentage participation of the SNI (Level) CONEVAL Academic Researcher Candidate, according to level within the System
3. National Researchers (Candidate, I, II, III).
3. Percentage participation of the SNI (Level) CONEVAL Academic Researcher Candidate, according to place of origin.
4. Identification of Word Cloud presenting the most important concepts expressed in the opinion of 104 SNI researchers specialized in Social Development.
5. Identification of the total number of identified clusters of 104 SIN researchers (Candidate, I, II and III) participating in the selection process of Academic Researchers for CONEVAL 2020.
6. Analysis of hierarchical grouping according to the theme identified in 104 SIN researchers (Candidate, I, II and III) participating in the selection process of CONEVAL 2020 Academic Researchers.

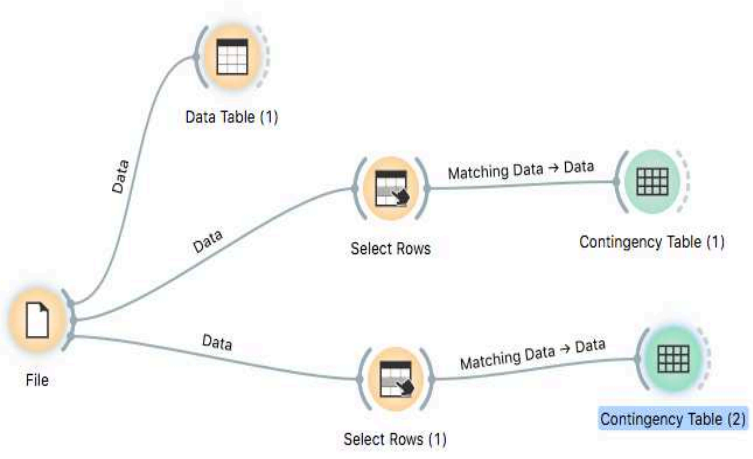
Methods

The information is presented in PDF (Portable Document Format), it corresponds to 104 personalized documents, each document consists of at least 4 pages in free format and in Spanish, the review and compliance with the regulations for public information indicated in the article. 113, fraction I of the Federal Law of Transparency and Access to Public Information, as well as the corresponding scan was carried out by the National Council for the Evaluation of Social Development Policy (CONEVAL), obtaining from the site: https://www.coneval.org.mx/quienessomos/InvestigadoresAcademicos/Paginas/Convocatoria_Eleccion_2020.aspx

The information in PDF documents show two characteristics: a) Legible for text mining treatment or b) Not legible for text mining treatment. For the PDF documents whose characteristics are "Not readable for text mining treatment", an Optical Character Recognition (OCR) process was carried out to modify them to a readable PDF format and prepare them for the use of the various mining modules. Data Mining software version 3.26.0.

The treatment of the information is carried out through the application software for text mining Orange version 3.26.0, in which the algorithm for the analysis of the analysis of participants in the Call for the election of three academic researchers to be part of the National Council for the Evaluation of Social Development Policy for CONEVAL 2020, dated January 9, 2020 (See Figure 1).

Figure 1. Algorithm for the analysis of participants in the Call for the election of three academic researchers to form part of the National Council for the Evaluation of Social Development Policy for CONEVAL 2020



Source: Self made. Orange Data Mining software version 3.26.0 text mining algorithm

In accordance with public access records at CONEVAL, the following information is available: (See Table 1 and 2, below).

1. A total of 104 participants belonging to the National System of Researchers were registered, of which 31.7% are women and 68.3 are men.

2. Of the 104 participating SNI researchers, 18.3% have a Candidate Level, 40.4% are Level I, 26.0% are Level II and 15.3% are Level III.
3. 52.9% of participating SNI researchers come from research institutions located in Mexico City

Table 1. Distribution according to sex and level in the National System of Investigators (SNI) of the participants in the selection process of Academic Researchers CONEVAL 2020

		Sexo		
		F	H	Σ
SNI	Candidato (a)	4	15	19
	SNI I	16	26	42
	SNI II	11	16	27
	SNI III	2	14	16
	Σ	33	71	104

Source: Own elaboration with information from the National Council for the Evaluation of Social Development Policy (CONEVAL) https://www.coneval.org.mx/quienessomos/InvestigadoresAcademicos/Paginas/Convocatoria_Eleccion_2020.aspx

Table 2. Distribution level in the National System of Investigators (SNI) and Federal Entity of origin of the participants in the selection process of Academic Researchers CONEVAL 2020

		Candidato (a)	SNI I	SNI II	SNI III	Σ
Estado	Aguascalientes	0	1	0	1	2
	Baja California	0	2	0	1	3
	Campeche	1	0	0	0	1
	Chiapas	0	1	1	0	2
	Ciudad de México	9	20	16	10	55
	Coahuila	0	2	0	0	2
	Estado de México	1	2	2	2	7
	Guanajuato	1	0	0	0	1
	Jalisco	0	3	1	1	5
	Michoacán	0	1	0	0	1
	Morelos	0	2	2	0	4
	Nayarit	1	0	0	0	1
	Nuevo León	0	2	2	0	4
	Oaxaca	0	1	0	0	1
	Puebla	1	4	2	0	7
	Quintana Roo	1	0	0	0	1
	Sonora	0	0	1	0	1
	Tabasco	1	0	0	1	2
	Veracruz	3	1	0	0	4
	Σ	19	42	27	16	104

Source: Own elaboration with information from the National Council for the Evaluation of Social Development Policy (CONEVAL) https://www.coneval.org.mx/quienessomos/InvestigadoresAcademicos/Paginas/Convocatoria_Eleccion_2020.aspx

The processing of the information is carried out through an algorithm for text mining to:

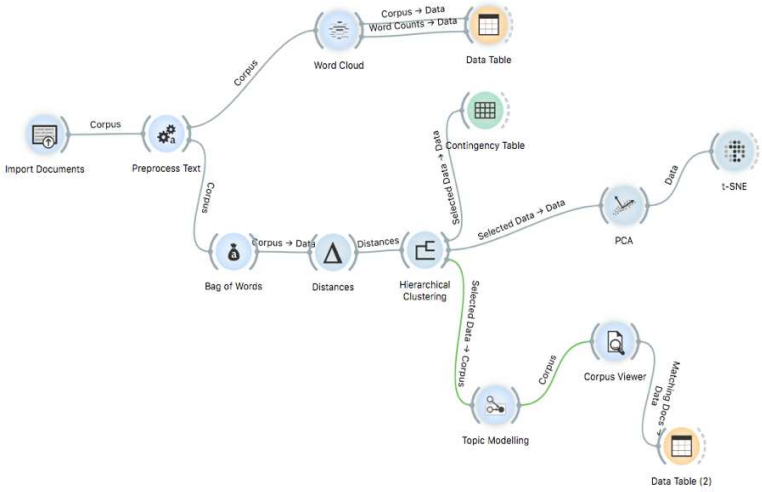
1. The construction of a word cloud that presents the most important concepts expressed in the opinion of 104 SNI researchers specialized in Social Development and that were presented to CONEVAL as part of the proposal for action and change to prepare scenarios for decision-making of social development policy in Mexico. (See Figure 2).

Figure 2. Word cloud presenting the most important concepts expressed in the opinion of 104 SNI researchers specialized in Social Development

Source: Own elaboration with information from the National Council for the Evaluation of Social Development Policy (CONEVAL)
https://www.coneval.org.mx/quienessomos/InvestigadoresAcademicos/Paginas/Convocatoria_Eleccion_2020.aspx

2. The cluster analysis. The algorithm presented allows us to identify similarities around the opinion of 104 SNI researchers specialized in Social Development and who were presented to CONEVAL as part of the proposal for action and change to prepare scenarios to make policy decisions. of social development in Mexico. (See Figure 3, below).

Figure 3. Text mining algorithm for cluster analysis identified as current challenges in the development of social policy and measurement in Mexico



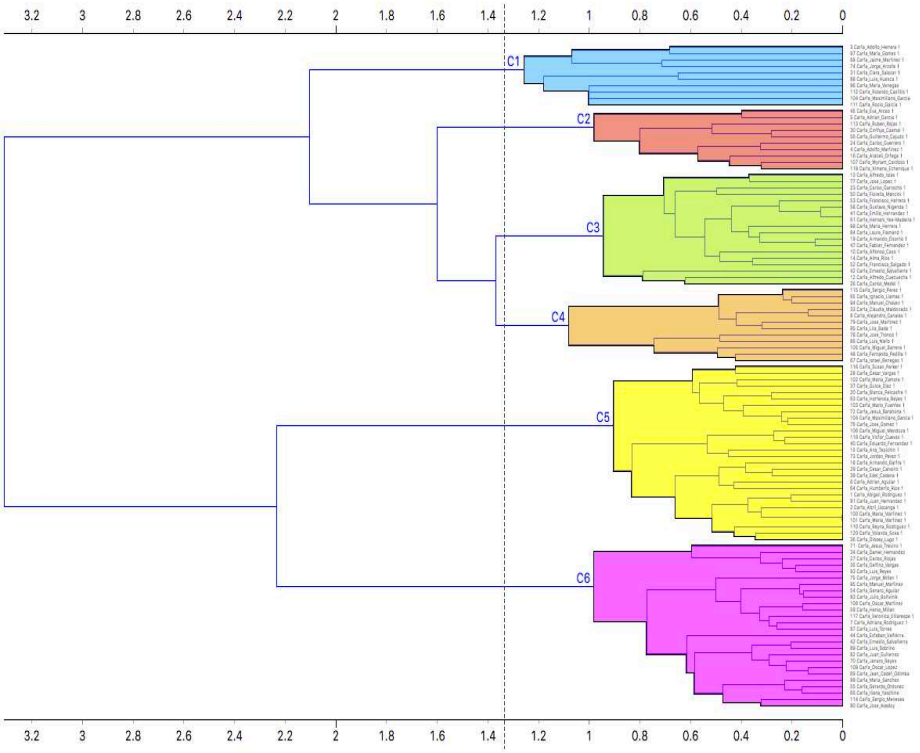
Source: Self made. Orange Data Mining software version 3.26.0 text mining algorithm

Table 3. Total number of identified clusters of 104 SIN researchers (Candidate, I, II and III) participating in the selection process of Academic Researchers for CONEVAL 2020

		Cluster						
		C1	**C2**	**C3**	**C4**	**C5**	**C6**	**Σ**
Cluster	**C1**	10	0	0	0	0	0	**10**
	C2	0	10	0	0	0	0	**10**
	C3	0	0	18	0	0	0	**18**
	C4	0	0	0	12	0	0	**12**
	C5	0	0	0	0	28	0	**28**
	C6	0	0	0	0	0	26	**26**
	Σ	**10**	**10**	**18**	**12**	**28**	**26**	**104**

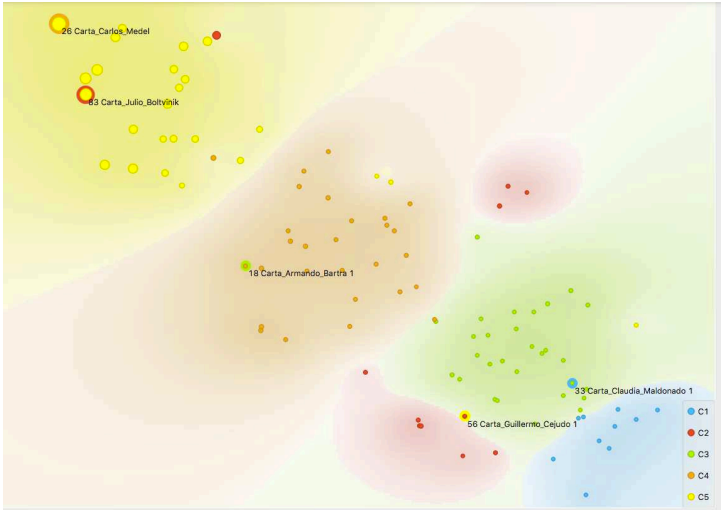
Source: Self made. Orange Data Mining software version 3.26.0 text mining algorithm

Figure 4. Analysis of hierarchical grouping according to the theme identified in 104 researchers SNI (Candidate, I, II and III) participants in the selection process of CONEVAL 2020 Academic Researchers



Source: Self made. Orange Data Mining software version 3.26.0 text mining algorithm

Figure 5. Analysis of hierarchical grouping according to the theme identified in 104 researchers SNI (Candidate, I, II and III) participants in the selection process of CONEVAL 2020 Academic Researchers



Source: Self made. Orange Data Mining software version 3.26.0 text mining algorithm

Declaration of competing interest

The authors declare that they do not have, or could be perceived to have, known competitive financial interests or personal relationships that have documented the work reported in this article.

References

[1] Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, Journal of Machine Learning Research 14(Aug): 2349−2353. https://dl.acm.org/doi/pdf/10.5555/2567709.2567736

[2] National Council for the Evaluation of Social Development Policy (CONEVAL) https://www.coneval.org.mx/quienessomos/InvestigadoresAcademicos/Paginas/Convocatoria_Eleccion_2020.aspx

[3] National Council for the Evaluation of Social Development Policy (CONEVAL). (2020). CALL for the election of three academic researchers to form part of the National Council for the Evaluation of Social Development Policy. Official Journal of the Federation. January 9, 2020 Recovered from: https://www.coneval.org.mx/quienessomos/InvestigadoresAcademicos/Documents/Diario_Oficial_de_la_Federacion.pdf

[4] National Council for the Evaluation of Social Development Policy (CONEVAL). (2020) Simplified privacy notice regarding the call for the election of three academic researchers to form part of the National Council for the Evaluation of Social Development Policy. Official Journal of the Federation. Recovered from: https://www.coneval.org.mx/odt/UAA/GCI/Documents/Avisos_Privacidad/Convocatoria_Consejeros/APS_Convocatoria_IA%20vf.pdf

[4] Software Orange Data Mining version 3.25.1 https://orange.biolab.si

[5] Software ocr.space version 3.50 https://ocr.space/

Reference to a dataset:

[6] Raw data can be retrieved from the Github repository https://github.com/CMedelR/Vertientes-propuestas-de-mejora-en-la-medicion-y-desarrollo-social/blob/master/README.md