

# SISTEMAS DE PENSIÓN LATINOAMERICANOS VISTOS DESDE LA ACADEMIA: MODELADO DE TÓPICOS A PARTIR DE TÍTULOS DE ARTÍCULOS CIENTÍFICOS

**María Fátima Dos Santos**

Universidad Central de Venezuela

maria.dossantos@ucv.edu.ve

2santos.mf@gmail.com

<https://orcid.org/0009-0007-7313-1705>

DOI: 10.37594/oratores.n21.1539

Fecha de recepción: 24/09/2024

Fecha de revisión: 05/11/2024

Fecha de aceptación: 05/12/2024

## RESUMEN

Durante los años '40 y '50 del siglo XX, la mayoría de los países latinoamericanos adoptaron Sistemas de Pensión de Vejez basados en la asistencia pública, que a finales de siglo presentaba dificultades propias de países en desarrollo (cobertura limitada, desigualdad, bajas tasas de reemplazo, etc.), lo cual propició la implementación de ajustes y reformas. A nivel mundial (incluyendo Latinoamérica), los Sistemas de Pensión enfrentan hoy en día numerosos desafíos, especialmente derivados del aumento de la longevidad y el descenso de la natalidad. Este artículo explora el discurso de la academia latinoamericana sobre el tema de los Sistemas de Pensión de Vejez en Latinoamérica, analizando un corpus de 317 títulos de artículos disponibles en el repositorio Scielo, mediante diferentes técnicas de Minería de Textos. Se utilizaron los módulos Text Mining (v3.1.11) y Text Table (v 1.16.1) de Orange Data Mining, mediante distintos procedimientos no supervisados (Word Cloud, Bag of Words, Extract Keyword) hasta llegar al Modelado de Tópicos con Alocaión Latente de Dirichlet. Luego de evaluar los indicadores cuantitativos y de explorar cualitativamente el contenido de los tópicos generados, se decidió elegir la solución de cuatro tópicos, los cuales podrían titularse respectivamente como “Económico-Sistémico” (características de los distintos Sistemas de Pensión de Vejez), “Beneficios” (servicios cubiertos por estos sistemas, como salud, alimentos, etc.), “Legal-Laboral” (aspectos legales y de derechos humanos) y “Acceso-Cobertura” (participación y desigualdad). Estos tópicos resumen los principales temas recurrentes en la discusión académica latinoamericana en torno a las Pensiones de Vejez.

**Palabras clave:** Sistemas de Pensión, Vejez, Latinoamérica, Redes Semánticas, Modelado de Tópicos, Inteligencia Artificial.

## **LATIN AMERICAN PENSION SYSTEMS AS SEEN FROM THE ACADEMY: MODELING TOPICS FROM SCIENTIFIC ARTICLE TITLES**

### **ABSTRACT**

During the 1940s and 1950s, most Latin American countries implemented pension systems based on public assistance, which at the end of the century presented difficulties typical of developing countries (limited coverage, inequality, low replacement rates, etc.), which led to the implementation of adjustments and reforms. Worldwide (including Latin America), Retirement Systems face numerous challenges today, especially those derived from the increase in longevity and the decrease in the birth rate. This article explores the discourse of Latin American academia on the subject of Retirement Systems in Latin America, analyzing a corpus of 317 titles of articles available in the Scielo repository, using different Text Mining techniques. The Text Mining (v3.1.11) and Text Table (v 1.16.1) modules of Orange Data Mining were used, through different unsupervised procedures (Word Cloud, Bag of Words, Extract Keyword) until reaching Topic Modeling with Dirichlet's Latent Allocation. After evaluating the quantitative indicators and exploring qualitatively the content of the topics generated, it was decided to choose the solution of four topics, which could be titled respectively as "Economic-Systemic" (characteristics of the different Old Age Pension Systems), "Benefits" (services covered by these systems, such as health, food, etc.), "Legal-Labor" (legal and human rights aspects) and "Access-Coverage" (participation and inequality). These topics summarize the main recurring themes in the Latin American academic discussion around Retirement Pensions.

**Keywords:** Retirement, Ageing, Latin America, Semantic Network, Topic Modelling, Artificial Intelligence.

### **Los Sistemas de Pensión en Latinoamérica**

La génesis de los Sistemas de Pensión se remonta a 1561, en Londres, con la Ley de Pobres. Alrededor de 1880, Nueva Zelanda, Alemania y Austria activan sistemas de ayuda a los necesitados, siendo la Alemania de Bismarck la primera en tener lo que podría llamarse con propiedad un sistema de Seguro Social para la vejez. En el siglo XX, numerosos países implementaron o ampliaron sistemas de apoyo social y vincularon los distintos subsistemas (vejez, incapacidad, etc). En 1919 se crea la Organización Internacional del Trabajo. (Cornejo, 2021)

En Latinoamérica, ya a principios del siglo XX algunos países poseían mecanismos de Seguridad Social, aunque limitados y desarticulados. En Argentina funcionaban fondos mutuales de contribución particular. En 1919, Uruguay legisla sobre el derecho a retiro del personal militar. En 1924, Chile implementa el Seguro Social obligatorio para accidentes laborales, enfermedad e

invalidez. En 1928 se crea el Instituto Ecuatoriano de Seguridad Social. En 1936, la reunión de la OIT se celebra en Chile. Ese mismo año se promulga la Ley del Seguro Social Obligatorio en Perú.

A nivel mundial, el tema de la Seguridad Social adquiere relevancia en la década de los 40, durante la guerra. (Cornejo, 2021). La mayoría de los países de Latinoamérica instauran sistemas de Seguridad Social articulados, creando organismos de coordinación nacional (“Institutos de Seguros Sociales”) que centralizan fondos e implementan políticas. Estos Institutos fueron fundados en 1941 en Panamá y Costa Rica, en 1943 en México y Paraguay, en 1944 en Venezuela, en 1945 en Colombia y en 1946 en Guatemala. Más tardíamente lo hicieron El Salvador (1949), Honduras (1954), Nicaragua (1955) y Bolivia (1956). En Cuba se instaura finalmente en 1963.

Debe recordarse que la Seguridad Social integra diferentes subsistemas, como jubilación (o Pensión de Vejez), desempleo, enfermedad, maternidad, viudez, etc. El foco de este artículo es el Subsistema (o Sistema) de Pensiones, dejando para futuros análisis los otros programas y sistemas de Seguridad Social. Pueden proponerse taxonomías para los Sistemas de Pensión de Vejez, a pesar de las profundas diferencias entre países. Por ejemplo, los Sistemas de Pensión mediterráneos (España, Portugal, Italia), resultan muy similares. También lo son los de los países anglosajones. Al final, las taxonomías reflejan similitudes entre países con historia, cultura y sistemas similares. (Witthehouse, 2007; Instituto Santa Lucía, sf). Así, los sistemas Latinoamericanos suelen ubicarse en un mismo grupo. Según Witthehouse (2007), en nuestro continente se observa un predominio de Sistemas Focalizados (las poblaciones más vulnerables tienen acceso a montos de pensión más altos) con Beneficios Definidos (el monto se ajusta dependiendo de los años trabajados y del salario devengado). Son sistemas de Contribución Definida (hay un aporte fijo mensual).

Los Sistemas de Pensión deben ser dinámicos, reformándose regularmente para adaptarse a los cambios sociales, históricos y económicos. Entre los últimos años del siglo pasado y los primeros del presente, varios países latinoamericanos hicieron modificaciones estructurales (cambios sistémicos) o paramétricas (adaptaciones menores, no sistémicas) (Mesa-Lago, 2009). En su mayoría, implicaron complementar el sistema público con alternativas privadas, aliviando la carga del Estado y (al menos teóricamente) mejorando los beneficios de los receptores.

### **Evaluación de los Sistemas de Pensión**

Ciertos indicadores reflejan la calidad de los Sistemas de Pensión. Según el Banco Mundial (2007), los criterios de evaluación pueden dividirse en primarios y secundarios. Los primarios son:

- a) Adecuación: el Sistema provee beneficios suficientes para los receptores, ayudándolos a prevenir o mitigar la pobreza. Incluye la “Tasa de Reemplazo”, que es el porcentaje del

salario que es conservado en la pensión.

- b) Accesibilidad: acceder a él está dentro de las posibilidades de los beneficiarios
- c) Sustentabilidad: tiene viabilidad financiera y es sostenible
- d) Igualdad: promueve la redistribución de recursos desde los más ricos a los más pobres y provee el mismo beneficio entre los diferentes grupos de ingresos
- e) Predictibilidad: provee beneficios que 1) son contemplados por ley y no dependen de la voluntad de los actores 2) tienen mecanismos que protegen contra la inflación y 3) no es sensible a los riesgos de la longevidad (en la medida de lo posible)
- f) Robustez: es resistente a shocks provenientes de la economía, la política o los cambios demográficos

Los criterios secundarios incluyen: minimizar las distorsiones del mercado laboral, capacidad de movilizar ahorros y contribución del sistema al desarrollo del mercado financiero

La CEPAL evalúa los sistemas según seis variables: cobertura, suficiencia y calidad (que incluye tasa de reemplazo), igual trato y solidaridad social, equidad de género, eficiencia (logros en función al monto invertido) y sustentabilidad (Mesa-Lago, 2009). En Latinoamérica se perfila la existencia de tres grupos de países, de acuerdo al desarrollo de sus sistema: los socialmente desarrollados en materia de Pensiones (Chile, Costa Rica, Uruguay, Brasil, Argentina y Panamá), los medianamente desarrollados (Colombia, Venezuela y México), y los poco desarrollados (Ecuador, El Salvador, Guatemala, República Dominicana, Perú, Bolivia, Nicaragua, Paraguay y Honduras) (Mesa-Lago, 2009).

### **Desafíos Actuales de los Sistemas de Pensión**

A nivel mundial, los Sistemas de Pensiones enfrentan presiones importantes. En los países desarrollados, el principal reto se deriva del doble efecto ejercido por el aumento de la longevidad y la reducción de la natalidad, lo cual hace que la base de cotizantes disminuya en comparación con el porcentaje de beneficiarios activos (Attanasio, Kitao y Violante, 2007).

En Latinoamérica el principal problema es la escasa cobertura de pensiones, con un promedio regional no ponderado de población cubierta de apenas 33%, que en algunos países desciende hasta apenas 12% (Mesa-Lago, 2009). Hay un elevado índice de pobreza, una alta incidencia de trabajo informal y una fuerte relación de dependencia (proporción de la población que no es económicamente activa, comparado con la proporción económicamente activa). (Instituto Santa Lucía, sf). También se observan inequidades en la cobertura, asociadas a la informalidad y a la pobreza, estando los más pobres, las mujeres y la población rural más desamparados. Ciertos

grupos laborales (militares, funcionarios públicos, congresistas, etc) acceden a mejores sistemas de pensión. (Mesa-Lago, 2009). Además, en Latinoamérica se esperan presiones provenientes de la reducción de la natalidad y el aumento de la longevidad, que son tendencias mundiales.

En este contexto, es interesante comprender qué revisiones y propuestas se están haciendo desde la academia en torno al tema de las Pensiones de Vejez en Latinoamérica. ¿De qué temas se está hablando? ¿Cuáles son las preocupaciones reconocidas por la comunidad científica? ¿A cuáles problemas se intenta dar respuesta?

Una alternativa para explorar esto es identificar tópicos en las publicaciones científicas latinoamericanas, para lo cual pueden ser provechosas las herramientas de Minería de Textos.

### **Minería de Textos y Modelado de Tópicos**

Los artículos científicos son producciones textuales, que, como todo texto, tienen tres niveles: fáctico (sintáctico), semántico y hermenéutico o interpretativo (Calventus, 2019).

A nivel fáctico (o sintáctico) se fragmentan los textos en unidades básicas, contando la frecuencia de aparición de palabras y frases. Los detractores de estos enfoques argumentan que la atomización lexicométrica destruye el corpus, haciéndolo irreconstruible. (Dos Santos, 2013).

El siguiente nivel (semántico) aborda la relación entre los símbolos y lo que ellos representan. El texto es procesado por segmentación, agrupación y categorización, para luego dar cuenta de las relaciones entre las categorías, buscando así extraer significado. Se critica que este procesamiento implica el riesgo de alterar el significado original, debido al peso de la interpretación del investigador (Bècue, Lebart y Rajadell, 1992; Moscoloni y Satriano, 2000).

El tercer nivel es el interpretativo-hermenéutico, donde se incorpora la pragmática, la intención. Según Calventus (2008), en este nivel se busca identificar los ejes temáticos, clarificar la estructura argumentativa en de ellos y analizar los efectos que se pretenden producir.

En el nivel sintáctico, es fácil entender las ventajas de la asistencia de computadoras. Hoy, con la Inteligencia Artificial (IA) y la potencia creciente del hardware, se puede explorar el nivel semántico (incluso el hermenéutico) en forma automatizada, aunque requiriendo la intervención del analista. Para ello hay desarrollos que van desde la Minería de Textos (más básico) hasta el Procesamiento Natural del Lenguaje (más propio de la IA).

En general, “*la minería de datos textuales se refiere al proceso de extraer patrones de conocimiento interesantes y no triviales a partir de documentos textuales*” (Gaikward, Chaugule y Patil, 2014, p 1). Se procede por etapas: extracción de información (segmentación, tokenización, etc.), categorización y clasificación (formación de categorías, grupos de palabras o de documentos), visualización y resumen (reducir la complejidad, entresacando el significado principal).

Para relacionar palabras y frases entre sí, un concepto central es el de Redes Semánticas. Según Doerfel (1998), en ellas veríamos a las palabras como unidades léxicas y a estas unidades como nodos interrelacionados. Dibujarlas mediante un mapa nos permite evidenciar las relaciones. El sentido del texto se revela a través de la forma de la red. La cantidad de inter-relaciones alrededor de una palabra y su distancia respecto a las demás palabras nos habla de su importancia. El análisis puede hacerse desde dos perspectivas (Doerfel, 1998): desde las palabras (la distancia y relación entre palabras revela la presencia de tópicos) o desde los sujetos (el análisis se basa en la ubicación de los sujetos sobre el mapa de palabras, lo cual puede revelar la existencia de grupos de hablantes). (Bècuc, Lebart, y Rajadell, 1992; Behar, 1993; Benzécri, 1984; Dos Santos, 2013)

Las computadoras hoy permiten lograr análisis más complejos. En los últimos años se ha desarrollado en forma importante el Procesamiento Natural del Lenguaje (Natural Language Processing, NLP), gracias al cual podemos interactuar con computadoras usando el lenguaje humano natural, en diferentes idiomas. Esto quiere decir que el usuario no necesita traducir los textos a códigos. El NLP es la base de la inteligencia artificial que nos permite hacerle preguntas y obtener respuestas de herramientas como ChatGPT, Bing o Alexa. (Joseph et al, 2016; Yao, 2017)

Un tipo de NLP es el NLU (Natural Language Understanding), que analiza textos reales. (Dahl, 2023; Khurana et al, 2023). Cuando se hace de modo “*No Supervisado*”, se parte del texto mismo para ir construyendo agregaciones, categorías y clasificaciones, sin que hayan instrucciones previas del analista ni una solución predefinida (Variable Dependiente). (Bail, 2016)

Con NLU se puede lograr un abordaje semántico y de hermenéutica computacionalmente asistida. Un procedimiento reciente es el Modelado de Tópicos, que Pykes (2023) define como:

(...) es un enfoque utilizado con frecuencia para descubrir patrones semánticos ocultos representados por un corpus de texto e identificar automáticamente los temas que existen dentro de él. Es decir, es un tipo de modelado estadístico, que aprovecha el aprendizaje automático no supervisado para analizar e identificar grupos o conjuntos de palabras similares dentro de un cuerpo de texto. (Pykes, 2023, sp)

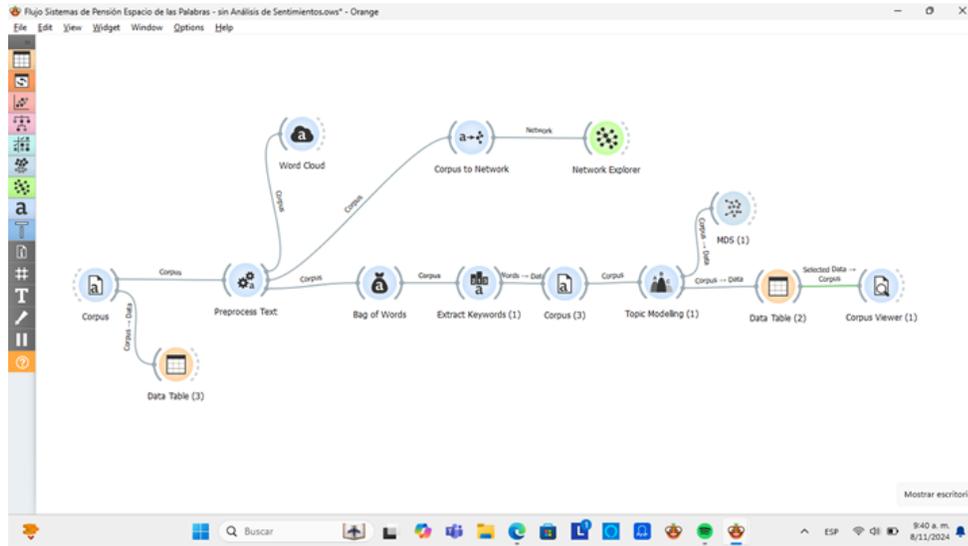
El Modelado de Tópicos procesa el texto como en la Minería de Textos, y luego, de forma automática, integra categorías para definir tópicos (Dahl, 2023; Jacobi et al, 2015). Incluso puede analizar el tono emocional predominante (Análisis de Sentimientos) (van Atteveldt et al, 2008). Según Pykes (2023): *“Esencialmente, el Modelado de Tópicos trabaja deduciendo palabras y agrupando las que son similares para formar clusters de tópicos”*.

Existen varias técnicas matemáticas de Modelado de Tópicos. Paranyushkin (2011) resalta cuatro, entre las cuales nos interesa especialmente la Asignación Latente de Dirichlet (LDA - Latent Dirichlet Allocation), que usa estadística bayesiana para asignar probabilidades a palabras, agrupándolas en clusters de significado. El proceso es inductivo y no supervisado y requiere la pericia del investigador con base en la teoría. Es decir, aunque el procesamiento es matemático, el análisis es interpretativo y cualitativo.

En este artículo se utilizó el sistema Orange Data Mining ([www.orangedatamining.com](http://www.orangedatamining.com)), que es un software estadístico no code de uso gratuito, programado a partir de Python, C++, C y otros lenguajes. Es desarrollado y mantenido por la Universidad de Liubliana y está disponible online. Está especialmente diseñado para visualización y análisis de data, Machine Learning y Data Mining. Aquí se utilizó la versión 3.37.0, que trae por defecto los módulos de análisis numérico. Los paquetes de Análisis de Textos y Redes que se usaron aquí deben ser descargados aparte (Text Mining versión 3.1.11, Text Table versión 1.16.1 y Network versión 1.9.0).

El programa tiene una interfaz gráfica muy amigable. Los procesos son cargados y estructurados sobre un lienzo (canva), tal como muestra la Figura 1, que representa los procesos usados en este artículo. Cada ícono es un paso del análisis. Algunos se refieren a la construcción de Corpus o Tablas, otros son análisis estadísticos (Bag of Words, Topic Modeling o Multidimensional Scaling MDS) y otros se refieren a graficación de resultados (Word Cloud, Network Explorer)

**Figura 1: Imagen del lienzo de Orange, que muestra los pasos del análisis presentado en este artículo**



**Fuente:** Elaboración propia

## METODOLOGÍA

Esta investigación explora un corpus de títulos de artículos científicos sobre el tema de los Sistemas de Pensión en Latinoamérica. El análisis se ha limitado a la base de datos del repositorio Scielo, que presenta buena cantidad de artículos provenientes de revistas latinoamericanas.

Las palabras clave usadas en la búsqueda fueron: “*PENSIÓN*” y “*VEJEZ*” o “*JUBILACIÓN*” (con esos conectores lógicos). Se limitó la búsqueda a las revistas publicadas en países latinoamericanos (Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Ecuador, México, Paraguay, Perú, Uruguay, Venezuela), excluyendo a los de habla portuguesa. Se excluye España, a pesar de compartir idioma, pues su Sistema de Pensiones es muy diferente y cae en otra taxonomía. El día 28 de octubre de 2024, la búsqueda arrojó un total de 391 artículos. La producción por período de publicación se muestra en la Tabla 1.

**Tabla 1: Número de artículos sobre pensiones compilados en Scielo por año de publicación**

Año de Publicación	N	%
De 2001 a 2005	12	3,1
De 2006 a 2010	52	13,3
De 2011 a 2015	93	23,8
De 2016 a 2020	152	38,9
2021 o posteriores	82	21,0

**Fuente:** Elaboración propia

La cantidad de artículos por período presenta una tendencia uniformemente creciente (el último de los rangos no ha terminado aún). No obstante, no puede saberse si se debe a un aumento en la relevancia del tema, o al crecimiento de Scielo por la incorporación de países y revistas.

Se limpió el corpus, eliminando títulos duplicados o en idiomas diferentes al español, conservando 345 artículos que se distribuyeron tal como se muestra en la Tabla 2.

**Tabla 2: Número de artículos sobre pensiones compilados en Scielo por país de publicación**

Artículos Publicados por País	N	%
México	114	25,3
Colombia	111	24,7
Chile	37	8,2
Perú	18	4,0
Argentina	17	3,8
Bolivia	12	2,7
Costa Rica	11	2,4
Cuba	8	1,8
Uruguay	6	1,3
Paraguay	4	0,9
Venezuela	4	0,9
Ecuador	3	0,7

**Fuente:** Elaboración propia

La mitad de las publicaciones fueron producidas en Colombia o México, seguido de Chile. Destaca, el bajo número de publicaciones de países que han tenido reformas en el sistema de pensión (como Argentina) o países con sistemas comprometidos (como Cuba y Venezuela).

Por último, se eliminaron títulos no relacionados con el tema (probables fallas del motor de búsqueda), conservando finalmente 317 artículos, corpus que se usará en las siguientes etapas.

## RESULTADOS

Sobre los títulos conservados se aplicaron técnicas de Minería de Textos (como Word Cloud, Bag of Words, Extract Keyword, etc), hasta llegar al Modelado de Tópicos. El tratamiento del corpus inicia con un preprocesamiento. Investigaciones recientes sugieren que aplicar una amplia batería de métodos de preprocesamiento permite lograr más ajuste y coherencia. (HaCohen-Kerner, Miller y Yigal, 2020). Las rutinas usadas en la presente investigación fueron:

- a) Transformación: Se cambió todo el texto a minúsculas y se removieron acentos.
- b) Tokenización: Se separaron las palabras usando el criterio de espacios en blanco. El análisis que aquí se presenta sólo utilizó unigramas (palabras simples).

- c) Normalización: Se reemplazan las palabras por su raíz. Por ejemplo, “estudiar”, “estudiante”, etc. se reemplazarían por la raíz “stud”. El método usado fue UDPipe Lematizer en Español (activada la casilla de UDPipe Tokenizer). A estas raíces se les denomina “tokens”. Luego se conservaron los 100 tokens más frecuentes.
- d) Filtrado: Se eliminaron números y palabras raras (conservando frecuencias relativas entre 0,1 y 0.9). Se eliminaron “stopwords” o “palabras vacías”, como preposiciones o artículos (y, a, o, la, etc). Se activaron POS tags y Regexp (que eliminan caracteres especiales).

Se conservaron palabras con frecuencia absoluta superior a 10 (según criterio empírico). Se solicitó una Nube de Palabras o Wordcloud, que se muestra en la Figura 2:

**Figura 2: Nube de Palabras de los títulos de artículos sobre Pensiones compilados en Scielo**

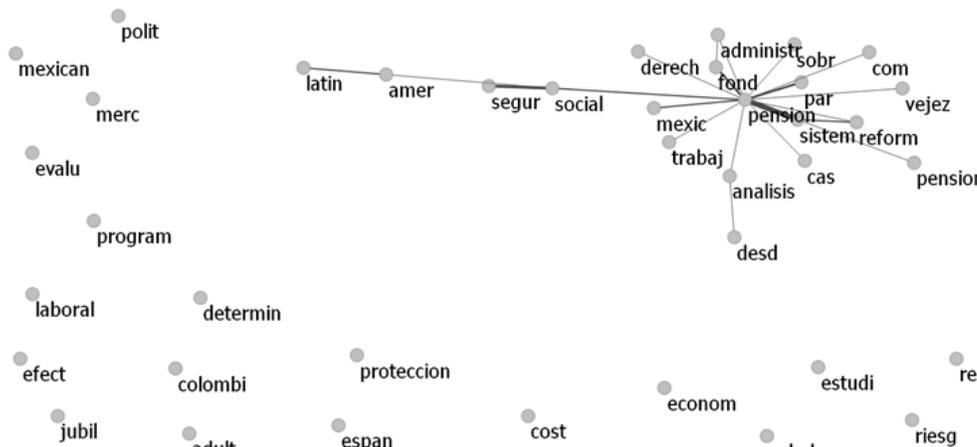


**Fuente:** Elaboración propia

Las palabras más frecuentes ocupan las posiciones centrales y son de mayor tamaño. Claramente son las palabras “*pensión*”, “*sistema*” y “*social*” (relacionándose con el término general de “*Sistemas de Seguridad Social*”). Son relevantes los tokens “*reform*” y “*analisis*”, sugiriendo temas sobre modificaciones a los sistemas (necesarias, propuestas o ejecutadas). Los tokens “*econom*”, “*administr*” y “*fond*”, sugieren preocupaciones económicas relacionadas con el sostenimiento del sistema. “*Protección*”, “*derech*”, “*program*”, “*laboral*” y “*trabaj*”, apuntan hacia una vertiente social y legal del tratamiento del problema. Los nombres de países (Colombia, México, Chile, Argentina) o nacionalidades (“*mexic*”), se asocian al origen de los artículos o de su ámbito de estudio.

La Nube de Palabras no refleja los vínculos entre ellas. La Exploración de Redes (Network Explorer) añade información al respecto, tal como se observa en la Figura 3:

**Figura 3: Red Semántica de los títulos de artículos sobre Pensiones compilados en Scielo**



**Fuente:** Elaboración propia

El mapa muestra las principales palabras del corpus y destaca su relación, cuando es relevante. El ancho de la línea que las vincula refleja la fuerza de esa relación. Vemos muchas palabras aisladas, indicando que no tienen un patrón claro de relación, aunque son frecuentes.

El punto más densamente conectado de la red tiene un centro formado por las raíces “*pension*” y “*sistem*”, relacionado también fuertemente con “*fond*” (de “*fondos*”). A la izquierda, “*segur*” y “*social*” están también muy vinculados y en el extremo es obvia la relación entre “*latin*” y “*amer*”. Algunas raíces aparecen claramente relacionadas con este núcleo, como “*derecho*”, “*reform*”, “*vejez*” y “*administr*”, indicando la importancia de los factores económicos y legales.

Se conservaron los 70 tokens más frecuentes y se corrió el procedimiento Bag of Words, donde se respetó la configuración por defecto de Orange. En este análisis, cada sujeto (artículo) es visto como una “*bolsa*” que contiene palabras o expresiones (conjuntos de dos o más palabras repetidas en varios textos, llamados bigramas, trigramas, etc). No se considera el orden de aparición, pero se mantiene la vinculación con el sujeto (artículo) que las emite. Bag of Words no muestra un resultado, pero permitirá hacer una clasificación automática de grupos de sentido.

Luego se extrajeron las palabras principales para conformar un nuevo corpus (ver pasos Extract Keywords y Corpus 3 en la Figura 1), usando el algoritmo YAKE!, un método autónomo (no depende de corpus externos o diccionarios), no supervisado y ligero (Sarwar, Noor y Miah, 2022).

YAKE! es actualmente un procedimiento popular para la extracción de palabras que funciona en diferentes idiomas, tipos y longitudes de texto. Muestra un coeficiente para cada palabra.

Finalmente, sobre este nuevo corpus se aplicó el Modelado de Tópicos (Topic Modelling), con Alocación Latente de Dirichlet, que sugiere un número óptimo de tópicos. En este caso, cuatro. Para evaluar la solución propuesta, se analizarán los valores de Perplejidad (Log Perplexity) y Coherencia (Topic Coherence). (HaCohen-Kerner, Miller y Yigal, 2020), que muestran la Tabla 3:

**Tabla 3: Indicadores de calidad de la solución con cuatro tópicos**

Log Perplexity	62.65430
Topic Coherence	0.85547

**Fuente:** Elaboración propia

La Coherencia habla de la homogeneidad interna (matemática) de los tópicos. El valor de la solución de cuatro tópicos (0,85547) es considerablemente alto (1 es el valor más alto posible).

La Perplejidad (Log Perplexity) es explicada como *“Un método estadístico utilizado para probar la eficiencia con la que un modelo puede manejar nuevos datos que no había visto antes”* (Hasan et al, 2021, p3). Es mejor una Perplejidad baja, siempre asociada a menor cantidad de tópicos. En este caso, el Log Perplexity es de 62.65430. La solución de tres tópicos tiene una Perplejidad menor (51.45112) y una Coherencia similar, pero el análisis cualitativo de las palabras contenidas en cada tópico nos hizo preferir la solución de cuatro.

La Tabla 4 muestra las palabras más importantes de cada tópico, con el coeficiente YAKE! palabra-tópico, seleccionadas según el principio de sedimentación (ordenadas decrecientemente, se muestran las más importantes hasta el momento en que el coeficiente exhibe una caída brusca).

**Tabla 4: Palabras más importantes en cada uno de los cuatro tópicos, ordenadas según YAKE!**

Palabras Importantes en el Tópico 1	YAKE!	Palabras Importantes en el Tópico 2	YAKE!	Palabras Importantes en el Tópico 3	YAKE!	Palabras Importantes en el Tópico 4	YAKE!
Colombia	0,624504	social	0,624732	personas	0,624507	acceso	0,624465
factores	0,624504	caso	0,624732	análisis	0,624504	enfoque	0,624464
jóvenes	0,624504	perspectivas	0,624422	derechos	0,624504	derecho	0,624463
económica	0,624504	alimentos	0,624421	jubilación	0,624504	propuesta	0,624463
adultos	0,624503	aproximación	0,624419	protección	0,624503	trabajo	0,624462
vejez	0,624503	riesgo	0,624418	fondos	0,624503	mirada	0,624462
sistemas	0,624503	invalidez	0,624418	Argentina	0,624502	estudio	0,624461
programa	0,624503	década	0,624417	laboral	0,624501	política	0,624461
relación	0,624503	constitucional	0,624416	evaluación	0,624498	Chile	0,624459
determinantes	0,624502	sostenibilidad	0,624415	años	0,624498	género	0,624459
pensiones	0,624501	trabajadoras	0,624415	mercado	0,624497	sistema	0,624459
seguridad	0,624501	mayores	0,624414	México	0,624497	características	0,624459
pensional	0,624499	periodo	0,624413	reforma	0,624497	salud	0,624458
mexicanos	0,624496	regulación	0,624413	envejecimiento	0,624494	colombiano	0,624457
		frontera	0,624413			multifondos	0,624456
		pensión	0,624412				

Fuente: Elaboración propia

Se han resaltado en amarillo las palabras más importantes para la comprensión de cada tópico, según un criterio cualitativo de significado. En definitiva, los cuatro tópicos podrían ser entonces descritos de la siguiente manera:

**TÓPICO 1: ECONÓMICO-SISTÉMICO.** Textos sobre la estructura de sistemas y programas de pensión, explorando tanto los elementos contributivos (por ello la presencia de palabras como “jóvenes” y “adultos”) como los asociados con la protección de la vejez. Destacan las palabras “Colombia” y “México”. Posiblemente este tema sea más frecuente en esos países debido a los procesos de reforma del sistema ocurridos durante los años reflejados en el corpus.

**TÓPICO 2: BENEFICIOS.** Aspectos relacionados con la extensión de los beneficios y la cobertura de los sistemas de Seguridad Social (por ello la presencia de palabras como “alimentos”, “invalidez”, “fronteras”, “trabajadoras”, “mayores”, etc), especialmente desde un enfoque legal (palabras como “regulación”, “constitucional”). Destaca la presencia de la palabra “sustentabilidad”, como expresión de la preocupación de que esta cobertura pueda ser respaldada.

**TÓPICO 3: LEGAL-LABORAL.** Temas relacionados con los derechos de los mayores

(“derechos”, “jubilación”, “protección”, “envejecimiento”, etc), aparentemente vinculado con el historial contributivo de los ahora pensionados (“personas”, “fondo”, “laboral”, “años”, “mercado”, etc). Este es el único tópico donde es importante la palabra “reforma”. Destacan México y Argentina, tal vez debido a los procesos de modificación del sistema en esos países.

**TÓPICO 4: ACCESO-COBERTURA.** Temas que discuten el acceso a los sistemas de pensión, con las palabras “acceso” y “género”, pero también “derecho”, “política” y “trabajo”. Este tópico analiza la cobertura en términos de población atendida (mientras que el Tópico 2 la aborda en función de los servicios cubiertos). Chile y Colombia aparecen destacados en este tema, y es el único donde encontramos el abordaje de un formato específico de Sistema de Pensión: Multifondos.

Estos resultados se revelan interesantes no sólo en términos de lo que muestran sino también en función de lo que no muestran, que podría dar pistas para áreas futuras de desarrollo:

- a) Los montos de cobertura (Tasa de Reemplazo) no parecen ser una preocupación, a pesar de que las pensiones en Latinoamérica son comparativamente bajas (Instituto Santa Lucía, sf), hasta el punto de ser insuficientes.
- b) Hay una escasa perspectiva de género, a pesar de que las mujeres en Latinoamérica tienen menos acceso al mercado laboral formal y a la prevalencia de la brecha salarial asociada al género. Sólo está tocado en el Tópico 2 (Beneficios) y no parece ser crucial.
- c) El aspecto de las necesarias reformas a los Sistemas de Pensión parece ser relevante sólo en el Tópico 3 (Legal-Laboral).
- d) La sustentabilidad de los sistemas sólo parece relevante en el Tópico 2, a pesar de que es un tema cardinal. Resulta notable que no haya sido una palabra importante en el Tópico 4 (Acceso-Cobertura).
- e) El cambio demográfico no parece ser relevante en ningún tópico.

## CONCLUSIONES

El análisis sugiere una creciente producción de artículos relacionados con el tema de las pensiones de vejez en Latinoamérica, impulsado fundamentalmente por las publicaciones de Colombia y México y, en menor medida, de Chile.

El análisis de frecuencia de las palabras, así como la exploración de la red semántica, muestra que el tema se asocia principalmente con aspectos sociales, económicos y legales. Vinculadas al núcleo de la red están palabras como pensión, vejez y laboral. Es notable que otras palabras frecuentes no presentan redes claras, lo cual podría sugerir debilidad de líneas temáticas (o del

relativamente bajo número de artículos disponibles). Palabras como “*protección*”, “*laboral*”, “*programa*” o “*política*” son frecuentes, pero no queda clara su articulación en la red semántica.

El Modelado de Tópicos sugiere la existencia de cuatro: Económico-Sistémico, Beneficios, Legal-Laboral y Acceso-Cobertura. Estos tópicos recuerdan los criterios de evaluación de Sistemas de Pensión, lo que sugiere que la academia aborda el tema sobre todo desde la evaluación de su eficiencia. Temas críticos parecen no estar reflejados (cambio demográfico y perspectiva de género, por ejemplo) o ser escasamente tocados (como la sustentabilidad). Los textos de los artículos académicos parecen preocuparse sobre todo de los derechos de los ciudadanos.

A nivel metodológico hay que decir que resultó difícil comprender las diferencias semánticas entre los tópicos. Los contenidos parecen entrelazarse, encontrando palabras similares en diferentes tópicos. Aunque esta dificultad es frecuente en la Minería de Textos, aquí podría haberse visto aumentada por la complejidad del tema de las pensiones o por el número relativamente bajo de artículos logrados.

También hay que destacar que el análisis omite una perspectiva cronológica que es relevante en este tema. Sería adecuado sondear los distintos temas a lo largo del tiempo, y vincular el análisis con los eventos históricos de cada país. Pero esto exigiría una data mucho más numerosa.

Sin embargo, el Modelado de Tópicos se revela como una herramienta rápida y poderosa para explorar temas en colecciones de datos textuales, incluyendo los académicos, siendo siempre necesario complementar el análisis con la interpretación humana de los resultados.

## REFERENCIAS BIBLIOGRÁFICAS

- Attanasio, O., Kitao, S., & Violante, G. L. (2007). Global demographic trends and social security reform. *Journal of Monetary Economics*, 54(1), 144-198. Recuperado de: <https://www.sciencedirect.com/science/article/pii/S0304393206002431>
- Bail, C. A. (2016). Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences of the United States of America*, 113(42), 11823–11828. <https://doi.org/10.1073/pnas.1607151113>
- Banco Mundial (2007). *The World Bank Pension Conceptual Framework*. Recuperado de <https://documents1.worldbank.org/curated/en/389011468314712045/pdf/457280BRI0Box31Concept1Sept20081pdf.pdf>
- Bècue, M., Lebart, L. y Rajadell, N. (1992). El análisis estadístico de datos textuales.

- La lectura según los escolares de enseñanza primaria [versión Adobe PDF]. Anuario de Psicología, (55), 7-22. Recuperado de <http://diposit.ub.edu/dspace/bitstream/2445/24200/1/79037.pdf>
- Behar, J. (1993). Aproximación al análisis textual informatizado, Anuario de Psicología, (59), 61-78. Recuperado de <http://diposit.ub.edu/dspace/bitstream/2445/24202/1/101702.pdf>
  - Benzécri, J.P. (1984). Description des textes et analyse documentaire. Les Cahiers de l'Analyse des Données, 9(2), 205-211. Recuperado de: [http://www.numdam.org/item/CAD\\_1984\\_\\_9\\_2\\_205\\_0.pdf](http://www.numdam.org/item/CAD_1984__9_2_205_0.pdf)
  - Calventus, J. (2008). Una aproximación al análisis de datos cualitativos textuales [Monografía]. Recuperado de <http://adcualimayor.googlepages.com/calventus2008.pdf>
  - Calventus, J. (2019). Análisis de datos textuales. Una primera aproximación. Revista Stultifera de Humanidades y Ciencias Sociales, 2(1), 50-62. <http://revistas.uach.cl/index.php/revstul/article/view/5825>
  - Cornejo, C.J. (2021). Análisis a los sistemas de pensiones del mundo a través de herramientas biplot para la clusterización mediante actores del sistema, indicadores del entorno y ranking mundial (usal.es). Recuperado de: <https://gredos.usal.es/handle/10366/149315>
  - Dahl, D. (2023) Natural Language Understanding with Python. Packt. Birmingham-Mumbai
  - Doerfel, M. L. (1998). What constitutes semantic network analysis? A comparison of research and methodologies. Connections, 21(2), 16-26. Recuperado de: <https://qualquant.org/wp-content/uploads/cda/Doerfel%20What%20constitutes%20semantic%20network%20analysis.pdf>
  - Dos Santos, M.F. (2013). Análisis Lexicométrico del Significado de la Palabra Dinero. Trabajo de Ascenso a Profesor Asistente. Universidad Central de Venezuela. Caracas.
  - Gaikwad, S.V., Chaugule, A., y Patil, P. (2014). Text mining methods and techniques. International Journal of Computer Applications, 85(17). Recuperado de 10.1.1.428.8805-libre.pdf
  - HaCohen-Kerner, Y., Miller, D., Yigal, Y. (2020) The influence of preprocessing on text classification using a bag-of-words representation. PLoS ONE 15(5): e0232525. Recuperado de <https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0232525&type=printable>
  - Hasan, M., Rahman, A., Karim, M. R., Khan, M. S. I., & Islam, M. J. (2021). Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA). In Proceedings of International Conference on Trends in Computational and Cognitive

- Engineering: Proceedings of TCCE 2020 (pp. 341-354). Springer Singapore. Recuperado de [Normalized-Approach-to-Find-Optimal-Number-of-Topics-in-Latent-Dirichlet-Allocation-LDA.pdf](#)
- Instituto Santa Lucía. (sf). Pensiones en Transición. Recuperado de [https://institutosantalucia.es/wp-content/uploads/2020/10/pensiones\\_en\\_transicion.pdf](https://institutosantalucia.es/wp-content/uploads/2020/10/pensiones_en_transicion.pdf)
  - Jacobi, C., Atteveldt, W. y Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4, 1-18. Recuperado de: <https://www.taylorfrancis.com/chapters/edit/10.4324/9781315115047-7/quantitative-analysis-large-amounts-journalistic-texts-using-topic-modelling-carina-jacobi-wouter-van-attveldt-kasper-welbers>
  - Joseph, S., Sedimo, K. y Kaniwa, F. y Hlomani, H. y Letsholo, K. (2016). Natural Language Processing: A Review, 6, 207-210. Recuperado de: <https://www.icts.res.in/sites/default/files/media/media-library/NLPIntro.pdf>
  - Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3), 3713–3744. <https://doi.org/10.1007/s11042-022-13428-4>
  - Mesa-Lago, C. (2009). Efectos de la crisis global sobre la seguridad social de salud y pensiones en América Latina y el Caribe y recomendaciones de políticas. Cepal. Recuperado de <https://repositorio.cepal.org/server/api/core/bitstreams/511cbdb3-54c7-4bde-8ed4-3234890b5116/content>
  - Moscoloni, N., y Satriano, C. (2000). Importancia del análisis textual como herramienta para el análisis del discurso. *Cinta de Moebio: Revista Electrónica de Epistemología de Ciencias Sociales*, (9), 4. Recuperado de [https://www.researchgate.net/profile/Nora-Moscoloni-3/publication/239567198\\_Importancia\\_del\\_Analisis\\_Textual\\_como\\_Herramienta\\_para\\_el\\_Analisis\\_del\\_Discurso/links/0c96053c7c524d03e5000000/Importancia-del-Analisis-Textual-como-Herramienta-para-el-Analisis-del-Discurso.pdf](https://www.researchgate.net/profile/Nora-Moscoloni-3/publication/239567198_Importancia_del_Analisis_Textual_como_Herramienta_para_el_Analisis_del_Discurso/links/0c96053c7c524d03e5000000/Importancia-del-Analisis-Textual-como-Herramienta-para-el-Analisis-del-Discurso.pdf)
  - Paranyushkin, D. (2011). Identifying the Pathways for Meaning Circulation using Text Network Analysis. *Nodus Labs*, 26, 1-26. Recuperado de: <https://noduslabs.com/wp-content/uploads/2012/04/Pathways-Meaning-Text-Network-Analysis.pdf>
  - Pykes, C. (2023). What is Topic Modeling? An Introduction With Examples. Recuperado de <https://www.datacamp.com/tutorial/what-is-topic-modeling>
  - Sarwar, T. B., Noor, N. M., & Miah, M. S. U. (2022). Evaluating keyphrase extraction algorithms for finding similar news articles using lexical similarity calculation and semantic relatedness measurement by word embedding. *PeerJ Computer Science*, 8, e1024. Recuperado de <https://peerj.com/articles/cs-1024.pdf>
  - van Atteveldt, W., Kleinnijenhuis, J. , Ruigrok, N. y Schlobach, S. (2008) Good News

- or Bad News? Conducting Sentiment Analysis on Dutch Text to Distinguish Between Positive and Negative Relations, *Journal of Information Technology & Politics*, 5(1), 73-94. Recuperado de: <https://www.tandfonline.com/doi/abs/10.1080/19331680802154145>
- Withehouse, Eduard (2007) *Pension Panorama: Retirement-income systems in 53 countries*. The World Bank. Whashington DF. Recuperado de: *Pensions Panorama* - ISBN: 0821367641 (uni-muenchen.de)
  - Yao, Mariya (2017). 4 Approaches To Natural Language Processing & Understanding. Recuperado de: <https://www.freecodecamp.org/news/how-natural-language-processing-powers-chatbots-4-common-approaches-a077a4de04d4/>