Análisis comparativo del contenido y temas legislativos sancionados durante los gobiernos de Iván Duque y Gustavo Petro en sus primeros dos años

[1]Laura Milena Londoño Navarro

Facultad de Derecho y Ciencias Políticas y Sociales

Maestría en Representación Política y Gestión Pública

Asesor (Alfredo Fernández Ortega)

28 de marzo de 2025

[1] Estudiante de la Universidad La Gran Colombia - correo: laura.londono0906@gmail.com

Resumen

Este estudio presenta un análisis comparativo del contenido legislativo sancionado durante los primeros dos años de los gobiernos de Iván Duque y Gustavo Petro en Colombia. Para ello, se emplearon diversas técnicas de extracción de texto (PyPDF y OCR Tesseract) que permitieron obtener de manera eficiente y precisa el contenido de las leyes. Posteriormente, se aplicaron metodologías avanzadas de Procesamiento de Lenguaje Natural (PLN) para el modelado y agrupación de tópicos utilizando la librería **BERTopic**. Este modelo, combinado con **UMAP** como técnica de reducción de dimensionalidad y **HDBSCAN** para el clustering, se configuró en tres métodos distintos que variaron en el modelo de embeddings, la vectorización del texto (completo vs. palabras clave extraídas con YAKE) y los parámetros de UMAP/HDBSCAN. El objetivo principal fue identificar y comparar las prioridades legislativas de ambos gobiernos mediante la frecuencia de palabras clave y la distribución temática de los documentos legislativos, aplicando filtros para eliminar términos irrelevantes y enfocándose en los términos más representativos de cada corpus.

Los resultados revelan diferencias significativas en las agendas legislativas de ambos periodos. La producción legislativa del gobierno de Gustavo Petro se centra en la protección de derechos (especialmente de mujeres y víctimas), la cooperación internacional, la promoción cultural y la conservación ambiental. En contraste, la producción legislativa del gobierno de Iván Duque prioriza la salud, la educación, las políticas financieras y la estabilización económica, con un enfoque adicional en el deporte y la seguridad alimentaria. Además, se observó que el método de Embeddings Robustos + Texto Completo, logró el mejor equilibrio entre diversidad temática y cohesión, asignando casi todos los documentos a tópicos coherentes. Este estudio subraya la eficacia de las técnicas de PLN y los modelos de embeddings en el análisis legislativo, proporcionando una herramienta valiosa para comprender las prioridades políticas y la evolución de las agendas públicas en diferentes administraciones.

Palabras Clave: *Análisis Legislativo, Procesamiento de Lenguaje Natural, OCR Tesseract, PyPDF, YAKE, BERTopic, UMAP, HDBSCAN, Gobierno Iván Duque, Gobierno Gustavo Petro, Embeddings en Español.*

Abstract

This study presents a comparative analysis of the legislative content enacted during the first two years of the administrations of Iván Duque and Gustavo Petro in Colombia. To achieve this, various text extraction techniques (PyPDF and OCR Tesseract) were employed to efficiently and accurately obtain the content of the laws. Subsequently, advanced Natural Language Processing (NLP) methodologies were applied for topic modeling and clustering using the **BERTopic** library. This model, combined with **UMAP** for dimensionality reduction and **HDBSCAN** for clustering, was configured in three distinct methods that varied in embedding models, text vectorization (full text vs. keyword extraction with YAKE), and UMAP/HDBSCAN parameters. The primary objective was to identify and compare the legislative priorities of both governments through keyword frequency and the thematic distribution of legislative documents, applying filters to eliminate irrelevant terms and focusing on the most representative terms in each legislative corpus.

The results reveal significant differences in the legislative agendas of both periods. Gustavo Petro's administration focuses on the protection of rights (especially for women and victims), international cooperation, cultural promotion, and environmental conservation. In contrast, Iván Duque's administration prioritizes health, education, financial policies, and economic stabilization, with additional emphasis on sports and food security. Furthermore, it was observed that **a** Robust Embeddings + Full Text Method achieved the best balance between thematic diversity and cohesion, assigning nearly all documents to coherent topics. This study underscores the effectiveness of NLP techniques and embedding models in legislative analysis, providing a valuable tool for understanding political priorities and the evolution of public agendas across different administrations.

Keywords: *Legislative Analysis, Natural Language Processing, OCR Tesseract, PyPDF, YAKE, BERTopic, UMAP, HDBSCAN, Iván Duque government, Gustavo Petro government, Embeddings.*

Introducción

El uso de técnicas de Procesamiento de Lenguaje Natural (NLP) resulta fundamental para manejar y analizar grandes volúmenes de documentos legislativos, los cuales suelen presentar una estructura compleja y un lenguaje especializado. Para llevar a cabo el estudio comparativo entre los primeros dos años de los gobiernos de Iván Duque y Gustavo Petro en Colombia, se emplearon métodos de extracción de texto (PyPDF y OCR Tesseract) que permiten obtener el contenido legal de forma confiable. A partir de esta extracción, se aplicó el algoritmo YAKE para la identificación de palabras clave, filtrando vocablos irrelevantes o demasiado frecuentes en el ámbito jurídico (e.g., "ley", "artículo", "nacional", "gobierno"). Dicho procedimiento proporcionó un listado preliminar de términos que sirvió como base para el posterior análisis temático.

Una vez obtenidas las palabras clave y el texto procesado, se utilizaron **modelos de *embeddings*** —entre ellos, **SentenceTransformer**, *all-mpnet-base-v2*, y *mrm8488/multilingual-e5-large-ft-sts-spanish*— para **vectorizar** el contenido legislativo de manera precisa y eficiente. Estos *embeddings* transforman cada documento (o fragmento de texto) en **representaciones numéricas** que capturan la semántica subyacente, permitiendo realizar **operaciones matemáticas** como el cálculo de distancias, la agrupación por similitud y la detección de patrones en espacios de alta dimensión. El uso de estos vectores constituye la pieza central de muchas aplicaciones de NLP, pues posibilita la comparación entre documentos por contexto y no solo por coincidencia de palabras exactas. Con dichas representaciones, la librería **BERTopic**, un modelo de lenguaje grande capaz de interpretar y clasificar texto, en conjunto con **UMAP** (técnica de reducción de dimensionalidad) y **HDBSCAN** (algoritmo de clustering basado en densidad), proveen un **modelado y agrupación de tópicos** robusto, revelando temas predominantes y relaciones temáticas que arrojan luz sobre la orientación de las políticas públicas en ambos periodos de gobierno.

Revisión de Literatura

Análisis Legislativo y su Importancia en la Ciencia Política

El análisis legislativo es una herramienta fundamental para comprender las prioridades y enfoques de las administraciones gubernamentales. A través del estudio de las leyes sancionadas, es posible identificar las áreas de interés, las políticas públicas promovidas y los cambios estructurales implementados en un periodo determinado. Este análisis no solo refleja la agenda política, sino que también permite evaluar el impacto de dichas leyes en la sociedad y en el desarrollo del país [1].

Procesamiento de Lenguaje Natural (PLN) en el Análisis de Textos Legislativos

El Procesamiento de Lenguaje Natural (PLN) es una rama de la inteligencia artificial que se enfoca en la interacción entre computadoras y lenguaje humano. En el contexto del análisis legislativo, el PLN facilita la extracción, interpretación y clasificación automática de grandes volúmenes de texto, permitiendo identificar patrones, temas y tendencias de manera eficiente [2].

Extracción de Texto de Documentos PDF

La mayoría de los documentos legislativos están disponibles en formato PDF, lo que presenta desafíos para su análisis automático debido a la variedad de formatos y la posible presencia de texto escaneado como imágenes. PyPDF y OCR Tesseract son dos herramientas ampliamente utilizadas para la extracción de texto:

- PyPDF: Es una librería de Python que permite la manipulación y extracción de texto de archivos PDF que contienen texto incrustado. Es eficiente pero puede enfrentar dificultades con PDFs escaneados o con estructuras complejas [3].

- OCR Tesseract: Es un motor de reconocimiento óptico de caracteres (OCR) que convierte imágenes de texto en texto editable. Es especialmente útil para PDFs escaneados o aquellos con tipografías no estándar, aunque puede requerir mayor capacidad de procesamiento [4].

Identificación de Palabras Clave con YAKE

YAKE (*Yet Another Keyword Extractor*) es un algoritmo no supervisado diseñado para la extracción de palabras clave de manera eficiente. Utiliza características como la frecuencia, la posición y la coocurrencia de términos dentro del texto para identificar las palabras más representativas de un documento. Este método es útil para reducir la dimensionalidad del texto y enfocarse en los términos más relevantes para el análisis posterior [5].

Modelado y Agrupación de Tópicos con BERTopic

BERTopic es una herramienta de modelado de tópicos que combina técnicas de aprendizaje profundo con algoritmos de clustering para identificar y agrupar temas en grandes conjuntos de datos textuales. Utiliza embeddings, que son representaciones vectoriales de texto generadas por modelos de lenguaje como SentenceTransformer, para capturar la semántica subyacente de los documentos. Posteriormente, aplica UMAP para la reducción de dimensionalidad y HDBSCAN para la agrupación de tópicos, permitiendo así una clasificación eficiente y coherente de los textos [6].

- UMAP (*Uniform Manifold Approximation and Projection*) es una técnica de reducción de dimensionalidad que preserva las relaciones locales entre los datos, facilitando la visualización y el clustering en espacios de menor dimensión [7].
- HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) es un algoritmo de clustering basado en densidad que identifica grupos de puntos densamente

conectados, permitiendo la detección de tópicos sin necesidad de predefinir el número de clusters [8].

Embeddings en PLN y su Papel en el Análisis Semántico

Los embeddings son representaciones vectoriales de palabras, frases o documentos que capturan su significado semántico en espacios de alta dimensión. Modelos como SentenceTransformer, que implementan arquitecturas basadas en Transformers, permiten generar embeddings que reflejan las relaciones contextuales y semánticas entre diferentes fragmentos de texto [9]. Estos vectores son esenciales para tareas de similitud, clustering y clasificación en PLN, ya que permiten operaciones matemáticas eficientes que facilitan la identificación de patrones y temas recurrentes en grandes corpus textuales.

Revisión de Estudios Previos

Aunque el uso de PLN en el análisis legislativo es un campo en crecimiento, existen pocos estudios que hayan comparado de manera exhaustiva las prioridades legislativas de diferentes administraciones mediante técnicas avanzadas de modelado de tópicos. Estudios previos se han enfocado en análisis de tendencias legislativas a través de métodos más tradicionales, pero la integración de herramientas como BERTopic ofrece una nueva perspectiva que combina la profundidad semántica con la eficiencia computacional [10].

Justificación del Estudio

Este trabajo se justifica por la necesidad de contar con metodologías robustas y automatizadas para el análisis legislativo, que permitan una comprensión detallada y comparativa de las agendas políticas de diferentes gobiernos. La aplicación de técnicas avanzadas de PLN, como BERTopic, facilita no solo la

identificación de los temas más relevantes, sino también la detección de cambios y prioridades a lo largo del tiempo, aportando una valiosa herramienta para la ciencia política y la gestión pública [11].

Metodología

Recolección de Datos

El primer paso consistió en la conformación de un corpus representativo de la producción legislativa de los gobiernos de Iván Duque y Gustavo Petro durante sus respectivos primeros dos años de mandato. Para ello, se descargaron desde la **página oficial del Departamento Administrativo de la Presidencia de la República (DAPRE)** los archivos PDF de las leyes sancionadas en cada periodo. Dichos documentos fueron recopilados y organizados en carpetas separadas, según el gobierno al que pertenecían, garantizando así un control adecuado de la fuente de datos. Esta estrategia de recolección facilitó el posterior procesamiento en bloque y la trazabilidad de cada archivo. Además, se tomaron en cuenta metadatos de fecha de sanción y número de ley para corroborar que la extracción coincidiese estrictamente con los primeros dos años de cada administración. Al final de este proceso, se obtuvo un **conjunto de documentos** en formato PDF, que sirvió como base para los análisis descriptivos y de *modelado y agrupación de tópicos*.

Extracción de Texto

Dado que los PDFs podían tener características diversas —algunos presentaban texto incrustado mientras que otros estaban escaneados como imágenes—, se evaluaron dos enfoques de extracción:

1. ***PyPDF:*** Una librería nativa de Python diseñada para la lectura y manipulación de archivos PDF. PyPDF funciona bien cuando el PDF contiene texto incrustado y una estructura reconocible, permitiendo una extracción relativamente sencilla.
2. ***OCR Tesseract (Pytesseract):*** Una solución de reconocimiento óptico de caracteres capaz de interpretar PDFs escaneados o con tipografías más complejas. Aunque Tesseract puede

requerir mayores recursos de cómputo, ofrece la ventaja de convertir imágenes en texto editable con alta precisión.

Durante las pruebas iniciales, OCR Tesseract se destacó por producir una extracción con menor ruido y una necesidad de preprocesamiento reducida, sobre todo en documentos con sellos oficiales, firmas digitales o tipografías poco estándar. Por esta razón, sus resultados fueron privilegiados en la fase siguiente. PyPDF se mantuvo como una alternativa útil en casos donde los PDFs ya tenían texto embebido. En conjunto, esta estrategia híbrida aseguró la optimización de la precisión de los datos para el análisis posterior.

Identificación de Palabras Clave

Una vez extraído el texto, se empleó la librería YAKE (*Yet Another Keyword Extractor*) para determinar las palabras o frases más relevantes en cada documento legislativo. YAKE realiza un análisis de frecuencias y coocurrencias, asignando a cada término un puntaje de relevancia. Como parte del flujo de trabajo, se incorporó un filtro de palabras comunes (e.g., “ley”, “artículo”, “nacional”, “gobierno”) y *stop words* del español, de modo que sólo permanecieran los vocablos con verdadero valor analítico. Adicionalmente, se definió un umbral para descartar términos excesivamente repetitivos y se obtuvieron así las 15 palabras clave más representativas para cada corpus (Duque y Petro). Este procedimiento facilitó la detección temprana de ejes temáticos, al tiempo que mitigó la interferencia de términos jurídicos genéricos o rutinarios.

Análisis de Tópicos (BERTopic)

Con el corpus ya limpio y las palabras clave seleccionadas, se abordó el modelado y agrupación de tópicos mediante la librería BERTopic, un modelo de lenguaje grande capaz de interpretar y clasificar texto de manera flexible. BERTopic se basa en una combinación de técnicas:

1. ***Embeddings:*** Representaciones semánticas del texto empleando diversos modelos de SentenceTransformer (e.g., *all-mpnet-base-v2*, *mrm8488/multilingual-e5-large-ft-sts-spanish*). Estos *embeddings* transforman frases o documentos en vectores numéricos, permitiendo medir su similitud semántica.
2. ***Reducción Dimensional (UMAP):*** Proyección de datos de alta a baja dimensión, conservando relaciones de cercanía entre vectores. UMAP facilita la visualización y el posterior agrupamiento temático de grandes volúmenes de texto.
3. ***Clustering (HDBSCAN):*** Algoritmo jerárquico basado en densidad, que permite detectar tópicos sin requerir un número fijo de grupos. HDBSCAN es particularmente robusto ante datos dispersos y puede dejar ciertos documentos sin asignar a ningún tópico si no cumplen un umbral mínimo de similitud.

Para cada gobierno (Duque y Petro) se aplicaron tres configuraciones o métodos de análisis, variando el modelo de embeddings, la forma de vectorizar (texto completo vs. palabras clave YAKE) y los parámetros de UMAP/HDBSCAN (e.g., número de componentes, min_cluster_size, cluster_selection_epsilon). Estas variaciones permitieron comparar la diversidad temática, la cohesión de los grupos y la cantidad de documentos sin asignar. Con ello se evaluó la efectividad de cada configuración, proporcionando una visión integral de los ejes legislativos prioritarios en ambos periodos de gobierno.

Resultados

Identificación de Palabras Clave

Gobierno Gustavo Petro

Después de aplicar filtros de palabras comunes y *stopwords* de baja aportación contextual, emergieron como más destacadas las palabras **"cultural" (11), "protección" (9), "internacional" (8), "acuerdo" (7), "convenio" (6), "educación" (6), "mujeres" (5)** y **"social" (5)**. Estos resultados sugieren una **agenda legislativa diversa**, donde se abarcan aspectos culturales, la protección de derechos, la cooperación internacional y la equidad de género.

Gobierno Iván Duque

Las palabras clave más frecuentes incluyeron **"salud" (8), "social" (7), "educación" (7), "sistema" (6), "pública" (6), "fondo" (5), "estampilla" (5)** y **"municipios" (5)**, lo que indica prioridades en salud, bienestar social, fortalecimiento educativo y aspectos financieros (fondos, estampillas).

Análisis de Tópicos: Gobierno Gustavo Petro

Se emplearon **tres métodos** con BERTopic para modelar y agrupar los textos legislativos correspondientes al gobierno Petro, variando *embeddings*, vectorización y parámetros de UMAP/HDBSCAN.

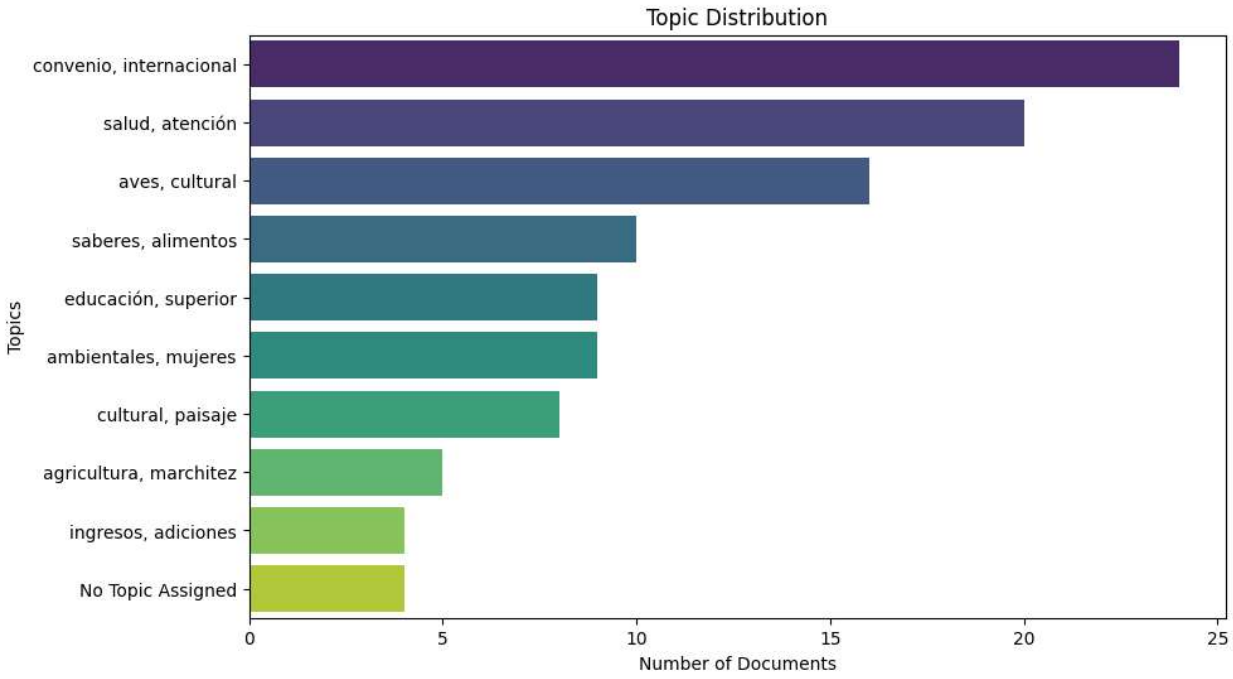
Figura 1. *Tópicos Gustavo Petro*

main_topic_label	count
convenio, internacional	24
salud, atención	20
aves, cultural	16
saberes, alimentos	10
educación, superior	9
ambientales, mujeres	9
cultural, paisaje	8
agricultura, marchitez	5
ingresos, adiciones	4
No Topic Assigned	4

dtype: int64

Elaboración propia.

Figura 2. *Frecuencia de Tópicos Gustavo Petro*



Elaboración propia.

Método 1: Embeddings + Texto Completo

- **Embeddings**: *mrm8488/multilingual-e5-large-ft-sts-spanish*
- **Reducción Dimensional (UMAP)**: 11 componentes, *n neighbors*=3
- **Clustering (HDBSCAN)**: *min_cluster_size*=3, *cluster_selection_epsilon*=0.01, *min_samples*=2
- **Vectorización**: *CountVectorizer* (stop words personalizadas, *min_df*=0.005, *max_df*=0.85)
- **Documentos sin asignar**: 2

Tabla 1. *Temas principales método 1 GP (Dos más representativos):*

Tópico Principal	N.º de Documentos
salud, ambientales	18
educación, superior	15

Elaboración propia.

Observación:

La alta cohesión (solo 2 documentos sin asignar) indica que las configuraciones elegidas son efectivas. Los temas prioritarios destacan la salud en relación con aspectos ambientales, y la educación superior.

Método 2: Embeddings + Palabras Clave (YAKE)

- **Embeddings**: *all-mpnet-base-v2*
- **Vectorización**: Palabras clave YAKE + *CountVectorizer*
- **Documentos sin asignar**: 14

Tabla 2. *Temas principales método 2 GP (Dos más representativos):*

Tópico Principal	N.º de Documentos
mujeres, atención	20
internacional, convenio	11

Elaboración propia.

Observación:

La especificidad temática es mayor (protección de mujeres, cooperación internacional), pero el número de documentos sin asignar aumentó a 14, lo que refleja una menor cohesión global al basarse únicamente en palabras clave.

Método 3: Embeddings Robustos + Texto Completo

- **Embeddings**: *mrm8488/multilingual-e5-large-ft-sts-spanish*
- **Reducción Dimensional (UMAP)**: 11 componentes, *n neighbors*=3
- **Clustering (HDBSCAN)**: *min_cluster_size*=3, *cluster_selection_epsilon*=0.01
- **Documentos sin asignar**: 2

Tabla 3. *Temas principales método 3 GP (Dos más representativos):*

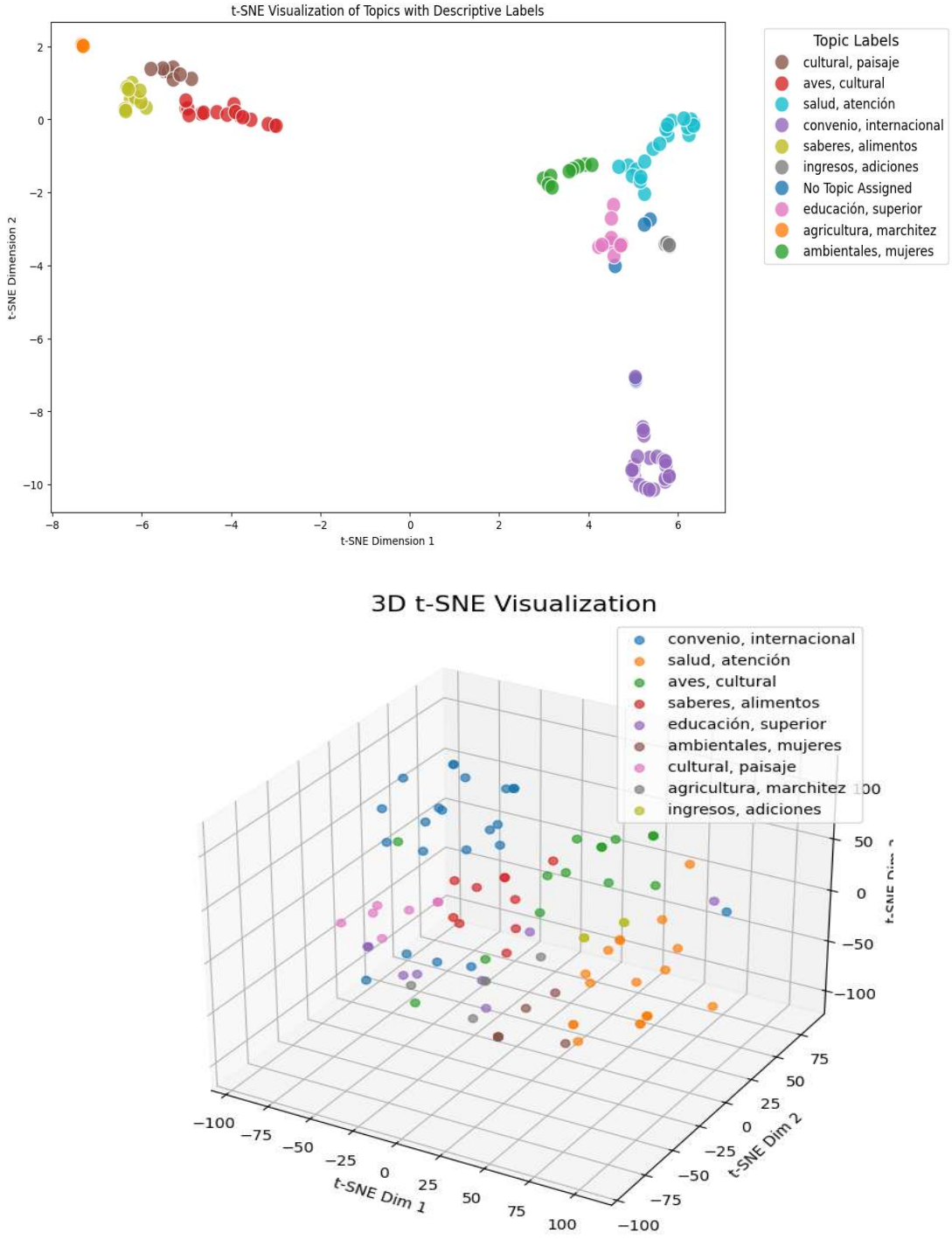
Tópico Principal	N.º de Documentos
convenio, naciones	19
ambientales, aves	17

Elaboración propia.

Observación:

El balance entre diversidad y cohesión es notable, asignando solo 2 documentos sin cluster. Se destacan acuerdos internacionales y asuntos ambientales, consolidando la idea de una agenda legislativa con foco en la cooperación y la protección de la fauna.

Figura 3. *Distribución espacial de tópicos GP*



Elaboración propia.

Figura 4. *Nube de palabras representativa para el tópico ambientales/mujeres - GP*

Elaboración propia.

Figura 5. *Nube de palabras del tópico saberes/alimentos - GP*

Elaboración propia.

Análisis de Tópicos: Gobierno Iván Duque

De igual forma, se aplicaron tres métodos a las leyes sancionadas durante el gobierno de Iván Duque. A continuación, se presentan los resultados obtenidos en cada uno de ellos.

Método 1: Embeddings + Texto Completo

- Embeddings: *all-mpnet-base-v2*
- UMAP: 6 componentes, *n_neighbors*=3
- HDBSCAN: *min_cluster_size*=3
- Documentos sin asignar: 12

Tabla 4. *Temas principales método 1 - IV (Dos más representativos):*

Tópico Principal	N.º de Documentos
deporte, salud	16
familias, informales	14

Elaboración propia.

Observación:

La atención recae en el binomio deporte-salud y en la protección de familias con empleos informales. Sin embargo, 12 documentos quedaron sin asignar, lo que sugiere una menor cohesión en la agrupación.

Método 2: Embeddings + Palabras Clave (YAKE)

- Embeddings: *all-mpnet-base-v2*
- Vectorización: Palabras clave YAKE
- Documentos sin asignar: 6

Tabla 5. *Temas principales método 2 - IV (Dos más representativos):*

Tópico Principal	N.º de Documentos
homenaje, fundación	18
estampilla, asamblea	18

Elaboración propia.

Observación:

La extracción de palabras clave mejoró la cohesión (solo 6 documentos fuera), identificando tópicos muy particulares como **homenajes** y creación de **estampillas** en asambleas.

Método 3: Embeddings Robustos + Texto Completo

- Embeddings: *mrm8488/multilingual-e5-large-ft-sts-spanish*
- UMAP: 11 componentes, *n neighbors*=3
- HDBSCAN: *min_cluster_size*=3, *cluster_selection_epsilon*=0.01
- Documentos sin asignar: 1

Tabla 6. *Temas principales método 3 - IV (Dos más representativos):*

Tópico Principal	N.º de Documentos
alimentos, fondo	22
estampilla, universidad	10

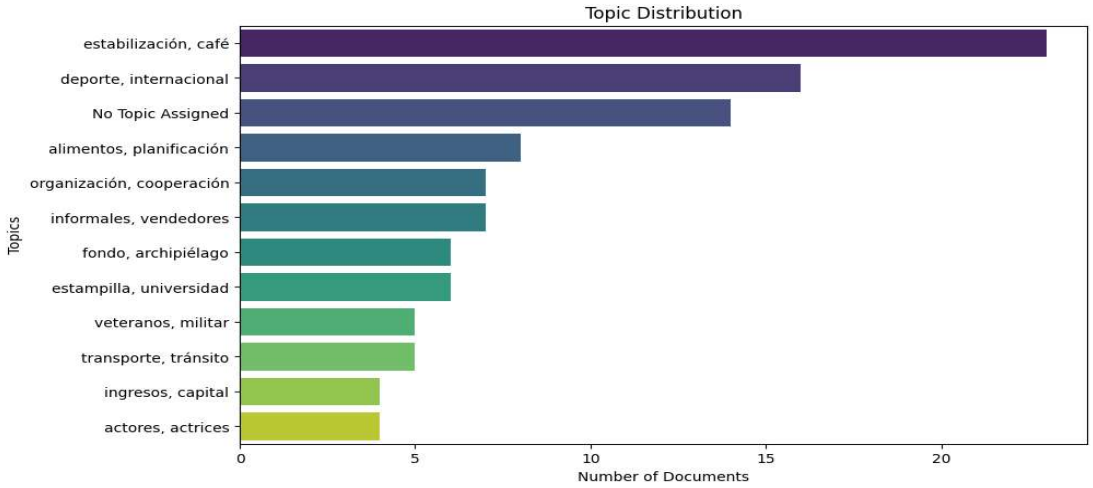
Elaboración propia.

Observación:

La adopción de embeddings multilingües avanzados redujo al mínimo (1) los documentos sin clasificar. Resaltan la seguridad alimentaria y la política fiscal como ámbitos prioritarios. Además, se identifican tópicos relacionados con la salud neonatal, historia clínica, y tránsito y asbesto, lo que denota una

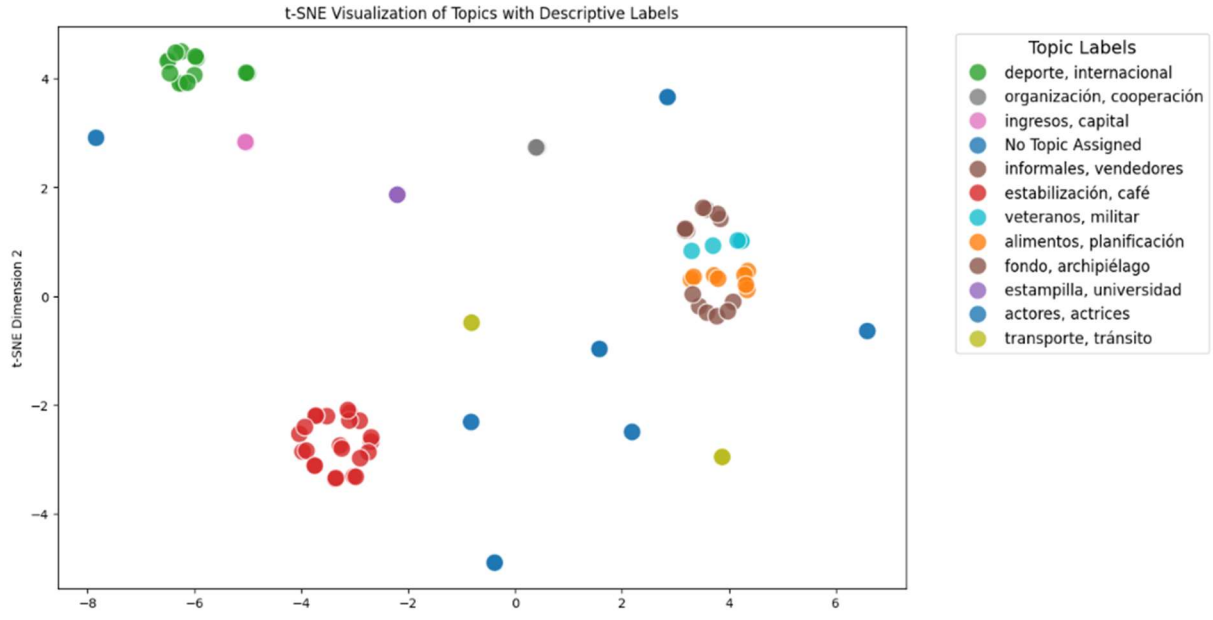
diversidad temática que abarca desde la regulación de salud y medio ambiente hasta la gestión fiscal y educativa. La mejora en la cohesión temática sugiere que este método captura de manera más eficiente la variedad de temas legislativos presentes en el corpus de Duque.

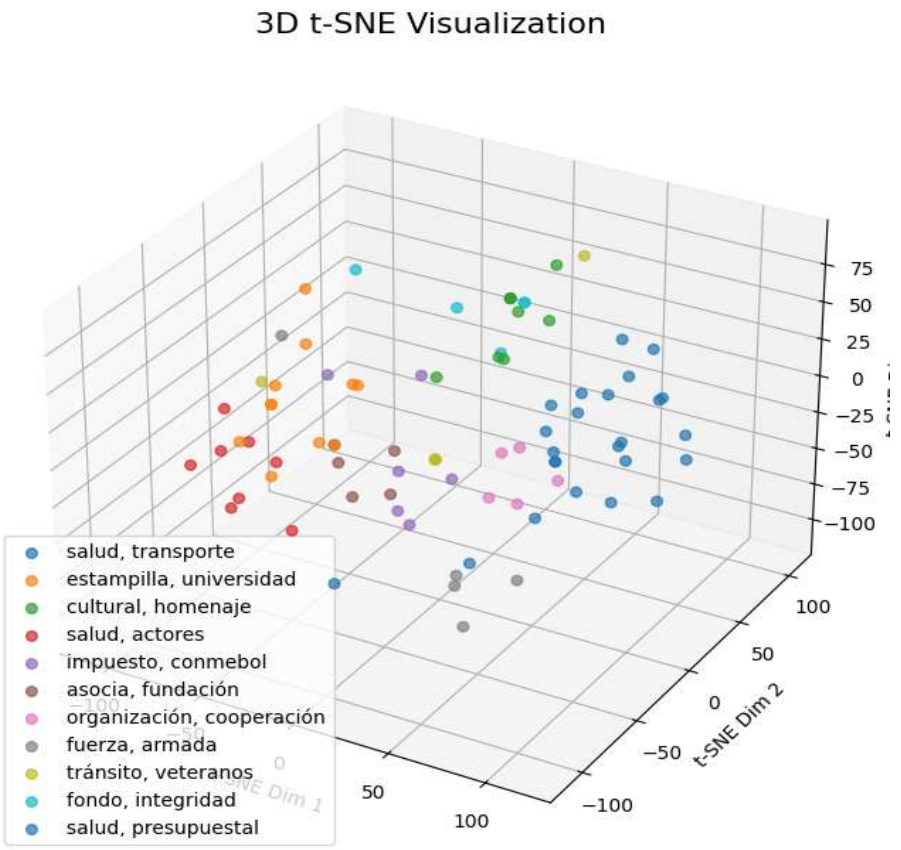
Figura 6. *Frecuencia de Tópicos Iván Duque*



Elaboración propia.

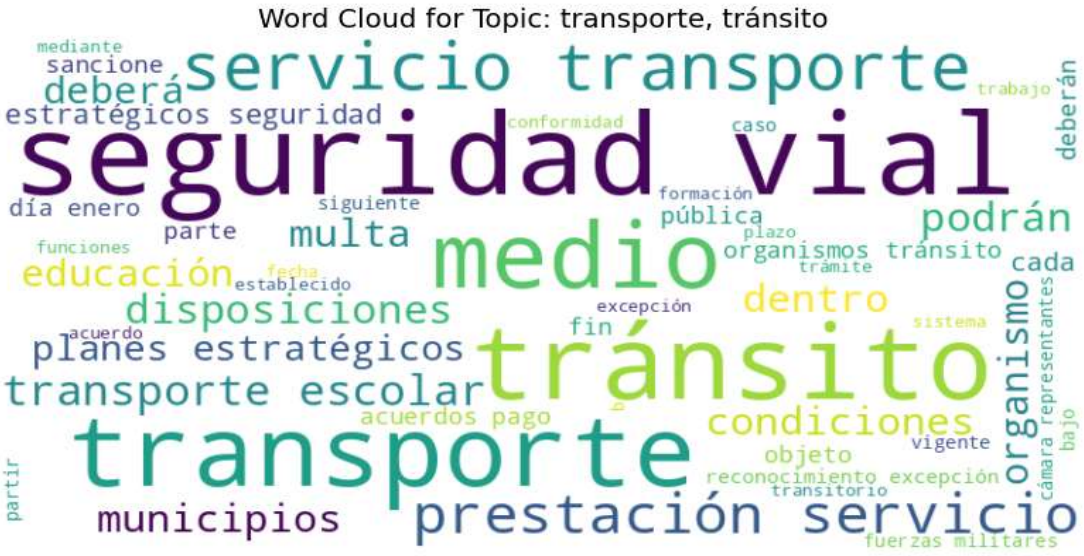
Figura 7. *Distribución espacial de tópicos GP*





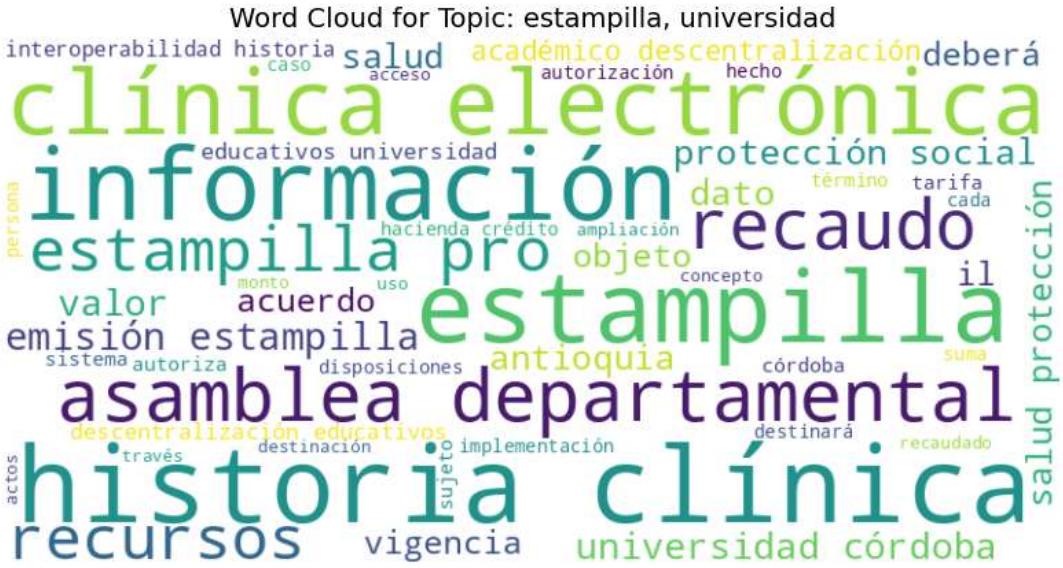
Elaboración propia.

Figura 7. *Nube de palabras representativas para el tópico transporte/tránsito - IV*



Elaboración propia.

Figura 7. *Nube de palabras representativas para el tópico estampilla/universidad - IV*



Elaboración propia.

Tabla 6. *Resumen de tablas 2.*

Gobierno / Método	Tópico 1	Docs	Tópico 2	Docs	Docs sin asignar
Petro - Método 1 (texto completo)	salud, ambientales	18	educación, superior	15	2
Petro - Método 2 (YAKE)	mujeres, atención	20	internacional, convenio	11	14
Petro - Método 3 (texto completo)	convenio, naciones	19	ambientales, aves	17	2
Duque - Método 1 (texto completo)	deporte, salud	16	familias, informales	14	12
Duque - Método 2 (YAKE)	homenaje, fundación	18	estampilla, asamblea	18	6
Duque - Método 3 (texto completo)	alimentos, derechos	21	impuesto, estabilización	14	1

Elaboración propia.

Este formato de tablas facilita la comparación entre métodos y gobiernos, mostrando de forma resumida los dos temas principales en cada análisis y el número de documentos agrupados en cada tópico. Además, se observa cómo la elección de *embeddings*, el uso de texto completo vs. palabras clave y los parámetros de UMAP/HDBSCAN inciden directamente en la cohesión temática y la cantidad de documentos sin asignar.

Tabla 7. *Resumen de tablas 1.*

GOBIERNO / MÉTODO	TÓPICO 1	TÓPICO 2	APLICACIÓN RECOMENDADA
Petro - Método 1 (Embeddings + Texto Completo)	salud, ambientales	educación, superior	Método 1:
Duque - Método 1 (Embeddings + Texto Completo)	deporte, salud	familias, informales	
Petro - Método 2 (Embeddings + Palabras Clave YAKE)	mujeres, atención	internacional, convenio	Método 2:
Duque - Método 2 (Embeddings + Palabras Clave YAKE)	homenaje, fundación	estampilla, asamblea	
Petro - Método 3 (Embeddings Robustos + Texto Completo)	convenio, naciones	ambientales, aves	**Método 3:**
Duque - Método 3 (Embeddings Robustos + Texto Completo)	alimentos, derechos	impuesto, estabilización	

Elaboración propia.

Discusión

Los resultados obtenidos tras comparar la producción legislativa de los gobiernos de Gustavo Petro e Iván Duque evidencian no solo las **diferencias en las prioridades políticas** de cada administración, sino también la **efectividad variable** de cada método de *modelado y agrupación de tópicos*. A continuación, se expone una discusión más profunda sobre la relevancia de los temas detectados, la coherencia de los métodos, y el **potencial de aplicación** de este enfoque para futuros análisis legislativos.

Diferencias Legislativas entre los Gobiernos

Gobierno Gustavo Petro

Protección de derechos y equidad: La alta frecuencia de tópicos enfocados en "mujeres, víctimas" y la reiterada aparición de "protección" sugieren un énfasis en políticas de género, asistencia social y defensa de grupos vulnerables.

Cooperación internacional: Palabras como "internacional", "convenio" y "acuerdo" reflejan la búsqueda de tratados o alianzas con organismos y naciones extranjeras, lo que sugiere una orientación más abierta a la diplomacia y a la negociación global.

Cultura y ambiente: El marcado interés en "cultural" y "ambientales" indica la intención de impulsar la identidad nacional, el patrimonio cultural y la conservación del ecosistema. Este hallazgo se ve reforzado por la presencia de términos específicos como "aves", lo que denota una política ambiental minuciosamente orientada a la biodiversidad.

Gobierno Iván Duque

Salud y educación: La predominancia de tópicos sobre "salud" y "educación" evidencia la búsqueda de reformas o fortalecimiento de sistemas públicos. La educación superior y la ampliación de coberturas sanitarias aparecen como pilares de varias de las leyes sancionadas.

Economía y financiamiento: Los términos "fondo", "estampilla" y "impuesto" apuntan a un interés en la creación o modificación de mecanismos de recaudación fiscal. Su repetición sugiere la priorización de la **estabilidad económica** y la financiación de proyectos, especialmente a nivel municipal y departamental.

Dimensión deportiva y social: El énfasis en "deporte, salud" y "familias informales" destaca iniciativas que buscan promover el bienestar físico y la inclusión social, posiblemente a través de políticas que impulsen la práctica deportiva y la formalización laboral.

En conjunto, ambos gobiernos comparten **ciertos ejes transversales** (salud, educación, atención social), pero divergen en **la orientación o profundidad** con que abordan estos temas, reflejando sus visiones específicas de política pública. Petro otorga un peso mayor a la cooperación internacional y a la protección de derechos y cultura, mientras que Duque prioriza el financiamiento, la estructura fiscal y la formalización económica.

Comparativa de Métodos de Modelado de Tópicos

Los tres métodos empleados (variando embeddings, vectorización y parámetros de reducción dimensional y clustering) ponen de manifiesto distintas **fortalezas y limitaciones**:

Método 1 (Embeddings + Texto Completo)

- **Ventaja**: Captura un panorama general, abarcando áreas amplias de la política pública.

- **Desventaja**: Tiende a dejar más documentos sin asignar cuando la diversidad temática es muy alta. Esto se evidenció especialmente en el gobierno Duque, con 12 documentos fuera de grupos.
- **Aplicación Recomendada**: Un primer cribado para obtener temas gruesos y pautar dónde profundizar.

Método 2 (Embeddings + Palabras Clave YAKE)

- **Ventaja**: Identifica tópicos específicos y de nicho, valiéndose de información muy concentrada.
- **Desventaja**: Al depender solo de palabras clave, sacrifica parte de la cohesión: el gobierno Petro presentó 14 documentos sin agrupar.
- **Aplicación Recomendada**: Análisis temáticos puntuales, donde se busquen rasgos distintivos o fenómenos emergentes.

Método 3 (Embeddings Robustos + Texto Completo)

- **Ventaja**: Ofrece un **equilibrio óptimo** entre diversidad y cohesión, asignando casi la totalidad de documentos en ambos gobiernos. Permite identificar tanto temas generales como específicos.
- **Desventaja**: Exige mayor capacidad de cómputo y un preprocesamiento más riguroso para vectorizar texto completo.
- **Aplicación Recomendada**: Estudios legislativos que requieran **gran precisión** y una visión integral de la agenda, idealmente cuando se dispone de recursos computacionales suficientes.

Consideraciones e Inferencias Propositivas

Profundidad Temática y Seguimiento de Implementación: El hecho de que “mujeres, víctimas” sea un tópico destacado en Petro sugiere no solo la aprobación de leyes sobre equidad de género y protección, sino también la necesidad de medir su **impacto real** y si hay recursos para su ejecución. Una extensión de este análisis podría ser la cuantificación del presupuesto asignado a dichas iniciativas.

Coherencia de la Agenda Internacional: Los tópicos “convenio, naciones” y “internacional, convenio” señalan un número significativo de leyes dedicadas a acuerdos bilaterales o multilaterales. Sería deseable explorar **el grado de ratificación** y la **implementación efectiva** de estos convenios, y si conectan con otras áreas prioritarias del plan de gobierno.

Reto de la Formalización Económica en Duque: La mención reiterada de “impuesto” y “estampilla” apuntala el interés en mecanismos de financiación y recaudación de fondos. Sin embargo, vale la pena investigar **cómo** estas iniciativas han afectado la economía local y si han logrado sus objetivos de estabilización o apoyo a sectores rurales (“campesina”).

Oportunidades de Mejora en la Cohesión Temática: Aunque el Método 3 mostró la mejor cohesión, aún se encontraron documentos sin asignar (especialmente con el gobierno Duque). Ajustar parámetros como *cluster_selection_epsilon* o *min_cluster_size*, o incluso combinar texto completo y palabras clave, podría mejorar la asignación en futuros análisis.

Finalmente, ambos gobiernos presentan **dos realidades legislativas** con solapamientos en educación y salud, pero divergentes en el grado de atención a convenios internacionales, cultura, género y finanzas. Esta comprensión más profunda de la agenda legislativa ofrece **lineamientos** para reformas, debates parlamentarios y mediciones de impacto político, haciendo uso de NLP y *embeddings* como herramientas de investigación y planificación.

Conclusiones y recomendaciones

El presente estudio ha logrado un análisis comparativo exhaustivo de las prioridades legislativas de los gobiernos de Gustavo Petro e Iván Duque durante sus primeros dos años de mandato en Colombia, utilizando técnicas avanzadas de Procesamiento de Lenguaje Natural (PLN). A continuación, se detallan las principales conclusiones derivadas de los hallazgos y su integración con la discusión previa:

Eficacia de las Técnicas de Extracción y Análisis de Texto

Extracción de Texto: OCR Tesseract vs. PyPDF

Los resultados demuestran que **OCR Tesseract** supera significativamente a **PyPDF** en términos de precisión y eficiencia, especialmente al procesar documentos PDF escaneados o con tipografías complejas. Esta superioridad se refleja en la menor cantidad de preprocesamiento requerido y en la mayor claridad de los textos extraídos, lo que facilita un análisis más fiable y detallado. Esta conclusión respalda la recomendación de utilizar OCR Tesseract como método preferido para la extracción de texto en futuros estudios legislativos que involucren documentos escaneados.

Identificación de Palabras Clave: Utilidad y Limitaciones de YAKE

La aplicación de **YAKE** (Yet Another Keyword Extractor) demostró ser una herramienta valiosa para la identificación preliminar de términos relevantes en los documentos legislativos. Sin embargo, se observó que depender exclusivamente de palabras clave puede llevar a una **subrepresentación del contexto completo** de cada ley, resultando en una mayor cantidad de documentos sin asignar. Esto sugiere que, si bien YAKE es útil para detectar temas específicos, es necesario complementarlo con métodos que consideren el texto completo para mantener una cohesión temática adecuada.

Evaluación de los Métodos de Modelado de Tópicos (BERTopic)

Método 1: Embeddings + Texto Completo

Este método proporcionó una **visión general sólida** de los temas legislativos, capturando áreas amplias como salud, educación y protección de derechos. No obstante, se identificó una **menor profundidad en ciertos tópicos** y una mayor proporción de documentos sin asignar en el caso del gobierno Duque. Esto indica que, aunque eficaz para agrupar textos con temáticas similares, puede ser insuficiente para manejar la diversidad temática sin ajustes adicionales.

Método 2: Embeddings + Palabras Clave (YAKE)

La combinación de embeddings con la vectorización basada en palabras clave permitió **identificar tópicos más específicos y de nicho**, como la protección animal y homenajes. Sin embargo, este enfoque sacrificó la **cohesión global**, dejando una cantidad considerable de documentos sin asignar, especialmente en el análisis de Petro. Esto subraya la necesidad de equilibrar la especificidad con la cobertura total del corpus legislativo.

Método 3: Embeddings Robustos + Texto Completo

Este método demostró ser el más **balanceado y efectivo**, logrando un **equilibrio óptimo entre diversidad y cohesión temática**. Al utilizar embeddings más robustos y analizar el texto completo, se asignaron casi todos los documentos a tópicos coherentes, reflejando fielmente las prioridades legislativas de ambos gobiernos. Este enfoque no solo capturó temas amplios sino también aspectos específicos, consolidando su utilidad para estudios legislativos que requieran una visión integral y detallada.

Estas variaciones permitieron comparar la **diversidad temática**, la **cohesión** de los grupos y la **cantidad de documentos sin asignar**. Con ello se evaluó la efectividad de cada configuración, proporcionando una visión integral de los ejes legislativos prioritarios en ambos periodos de gobierno.

Comparativa de Prioridades Legislativas entre los Gobiernos

Gobierno Gustavo Petro

Las leyes sancionadas durante el mandato de Gustavo Petro reflejan una agenda legislativa profundamente diversificada, con un énfasis particular en la protección de derechos fundamentales, especialmente aquellos relacionados con las mujeres y las víctimas de violencia. Este enfoque legislativo destaca una fuerte intención por parte del gobierno de priorizar la equidad de género y la protección de sectores históricamente marginados. Temas como "mujeres, participación" y "víctimas, derechos" se encuentran entre los tópicos más frecuentes, lo que evidencia un compromiso consistente con la creación de un marco normativo que promueva el empoderamiento femenino, la prevención de la violencia de género y el acceso equitativo a oportunidades. Adicionalmente, las iniciativas legislativas también se centraron en acuerdos internacionales, mostrando un interés por fortalecer la cooperación multilateral en áreas como la protección de derechos humanos, el desarrollo sostenible y la integración cultural.

En línea con esta visión inclusiva y sostenible, la producción legislativa del gobierno de Petro otorgó gran relevancia a temas culturales y ambientales. Tópicos como "cultural, inmaterial" y "ambientales, aves" ilustran el esfuerzo por proteger el patrimonio cultural y natural del país. Las políticas culturales apuntaron al fortalecimiento de la identidad nacional mediante la promoción de expresiones artísticas, tradiciones y paisajes emblemáticos. Por otro lado, el énfasis en la conservación ambiental refleja una postura alineada con los Objetivos de Desarrollo Sostenible, priorizando la biodiversidad y la mitigación del cambio climático. Este conjunto de prioridades legislativas subraya una orientación hacia una política pública que no solo busca atender necesidades inmediatas, sino también construir un legado a largo plazo que fomente la inclusión social, la justicia ambiental y la integración global.

Gobierno Iván Duque

El periodo legislativo bajo el mandato de Iván Duque se caracterizó por un enfoque marcado en áreas clave como la salud, la educación y la estabilidad económica. La salud pública, en particular, emergió como un eje central, con iniciativas destinadas a ampliar la cobertura, fortalecer los sistemas existentes y mejorar el acceso a servicios esenciales. Simultáneamente, la educación fue otra prioridad destacada, especialmente en la promoción de la educación superior mediante instrumentos como las estampillas universitarias, que facilitaron la recaudación de fondos para instituciones educativas. Estas acciones reflejan un compromiso por parte del gobierno con el desarrollo del capital humano como motor para el crecimiento social y económico del país. Asimismo, las políticas en torno a la estabilización económica evidencian un enfoque pragmático hacia la gestión fiscal, incluyendo temas relacionados con impuestos, fondos públicos y estrategias de financiamiento.

Adicionalmente, el gobierno de Duque dio importancia a la formalización laboral, la promoción del deporte y la seguridad alimentaria, temas que complementaron su agenda orientada al bienestar social. La presencia de tópicos como "deporte, salud" y "alimentos, planificación" refleja la intención de fomentar estilos de vida saludables y garantizar el acceso equitativo a alimentos básicos, promoviendo simultáneamente el desarrollo rural y la sostenibilidad. Por otro lado, la priorización de temas fiscales y administrativos, como los impuestos y las estampillas, subraya un enfoque en la eficiencia en la recaudación de fondos públicos para sostener programas sociales y proyectos de infraestructura. Esta combinación de prioridades evidencia una administración preocupada por equilibrar las necesidades inmediatas de los ciudadanos con la sostenibilidad financiera a largo plazo, reflejando un enfoque de política pública pragmático y estructurado.

Implicaciones y Aplicaciones Futuras

Profundización en el Impacto Legislativo: Los tópicos identificados, especialmente aquellos relacionados con la protección de derechos y la cooperación internacional, abren la posibilidad de realizar análisis de impacto sobre cómo estas leyes han influido en la sociedad y en la administración pública. Futuras investigaciones podrían enfocarse en la evaluación del cumplimiento y efectividad de dichas legislaciones, así como en la asignación presupuestaria destinada a su implementación.

Seguimiento de Políticas Públicas: La capacidad de BERTopic para identificar y agrupar tópicos legislativos de manera coherente y específica proporciona una herramienta poderosa para el seguimiento continuo de políticas públicas. Esta metodología puede ser aplicada para monitorear la evolución de la agenda legislativa a lo largo del tiempo, permitiendo detectar tendencias emergentes y cambios en las prioridades gubernamentales.

Comparaciones con Otros Periodos de Gobierno: Extender este análisis a otros periodos gubernamentales facilitaría una comparación longitudinal, permitiendo identificar patrones históricos y cambios en las prioridades políticas de Colombia. Esto contribuiría a una comprensión más profunda de la dinámica legislativa y su relación con los contextos políticos y económicos de cada periodo.

Optimización de Métodos de Análisis: Los hallazgos sugieren que la combinación de embeddings robustos y análisis de texto completo es la más efectiva para capturar la diversidad temática sin sacrificar la cohesión. Sin embargo, hay oportunidades para mejorar aún más estos métodos, como la optimización de parámetros de UMAP y HDBSCAN, o la integración de técnicas híbridas que combinen texto completo y palabras clave para maximizar la precisión y cobertura del análisis.

Conclusión General

Este estudio ha demostrado la importancia y potencia de las técnicas de PLN y los modelos de embeddings en el análisis legislativo, proporcionando una herramienta versátil para desentrañar las prioridades políticas y las tendencias legislativas. La comparativa entre los gobiernos de Gustavo Petro e Iván Duque no solo revela sus diferentes enfoques legislativos, sino que también subraya la efectividad de métodos avanzados como BERTopic para realizar análisis detallados y precisos. En definitiva, el uso de embeddings multilingües y técnicas de modelado de tópicos se consolida como una metodología esencial para estudios legislativos y políticos, facilitando una comprensión más rica y matizada de las decisiones legislativas y sus implicaciones sociales y económicas.

Lista de Referencias

Campos, R., Mangaravite, V., Pasquali, A., Jatowt, A., Jorge, A., Nunes, C. and Jatowt, A. (2020). YAKE! Keyword Extraction from Single Documents using Multiple Local Features. In Information Sciences Journal. Elsevier, Vol 509, pp 257-289. https://www.sciencedirect.com/science/article/abs/pii/S0020025519308588?via%3Dihub

Campello, R. J. G. B., Kröger, P., Sander, J., & Zimek, A. (2020). Density-based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(2), Article e1343. https://doi.org/10.1002/widm.1343. https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1343

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://arxiv.org/abs/2203.05794

Jurafsky, Daniel & Martin, James. (2008). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. https://arxiv.org/abs/1802.03426

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084. https://arxiv.org/abs/1908.10084

Pino Montoya, José Wilmar. (2017). Aspectos metodológicos para evaluar una política pública. Rev. Humanismo y Sociedad, 5(1), 1-7. https://doi.org/10.22209/rhs.v5n1a01

De Koning, B. (2022, July). Extracting sections from PDF-formatted CTI reports. Retrieved from http://essay.utwente.nl/91727/

Rosati, G. (2022). Procesamiento de lenguaje natural aplicado a las ciencias sociales: Detección de tópicos en letras de tango. Revista Latinoamericana de Metodología de la Investigación Social, 12(23), 38–60.https://ri.conicet.gov.ar/handle/11336/187219

Jiménez Jaimes, E. L. (2024). Análisis de discurso basado en modelos grandes de lenguaje [Master's thesis, Universidad EAFIT]. Universidad EAFIT Repository. https://repository.eafit.edu.co/server/api/core/bitstreams/26661e34-ba02-401b-8f95-b57cedb6c5a7/content

Corona-Bermúdez, M. C. (2024). Procesamiento del lenguaje natural para el modelado de tópicos en ideas de innovación. Trabajo de obtención de grado, Maestría en Ciencia de Datos. Tlaquepaque, Jalisco: ITESO.

Grisales-Aguirre, A. M., & Figueroa-Vallejo, C. J. (2022). Modelado de tópicos aplicado al análisis del papel del aprendizajeautomático en revisiones sistemáticas. Revista de Investigación, Desarrollo e Innovación, 12(2), 279-292. https://revistas.uptc.edu.co/index.php/investigacion_duitama/article/view/15271/12481

Barahona Pinilla, B. (2023). Herramienta de apoyo para la identificación y análisis de temáticas de textos a través del modelado de tópicos. Universidad de los Andes. Disponible en: http://hdl.handle.net/1992/68874

Matallana Villegas, S. (2023). El uso de la inteligencia artificial en el análisis de impacto normativo. IUS ET SCIENTIA, 9(1), 9–22. https://doi.org/10.12795/IESTSCIENTIA.2023.i01.02

Vannoni M, Ash E, Morelli M. Measuring Discretion and Delegation in Legislative Texts: Methods and Application to US States. Political Analysis. 2021;29(1):43-57. doi:10.1017/pan.2020.9

ANASTASOPOULOS LJ, BERTELLI AM. Understanding Delegation Through Machine Learning: A Method and Application to the European Union. American Political Science Review. 2020;114(1):291-301. doi:10.1017/S0003055419000522